DOCUMENT RESUME

ED 152 859　　95　　TM 007 218

AUTHOR	Rose, Clare; Nyre, Glenn F.
TITLE	The Practice of Evaluation. ERIC/TM Report 65.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	ERIC-TM-65
PUB DATE	Dec 77
NOTE	95p.
AVAILABLE FROM	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, New Jersey 08541 ($5.00)
EDRS PRICE	MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS	Bibliographic Citations; *Case Studies; *Curriculum Evaluation; Early Childhood Education; Elementary Secondary Education; Evaluation; *Evaluation Methods; Evaluators; Higher Education; *Models; *Needs Assessment; *Program Evaluation; Research Design; Research Methodology; Research Utilization; State of the Art Reviews; Theories
IDENTIFIERS	Information Analysis Products

ABSTRACT

The first half of this monograph provides an overview of the theoretical concerns of evaluators. Definitions are provided of accountability, measurement, assessment, evaluation research, formative and summative evaluation, goal-free evaluation, goal-based evaluation, and evaluation. Several models of evaluation are described and discussed, including the Countenance Model, by Robert Stake; several Goal Attainment Models; the Discrepancy Model by Malcolm Provus; the CIPP (context, input, process, product) model by Egon Guba and Daniel Stufflebeam; and the decision-oriented model developed at UCLA's Center for the Study of Evaluation by Marvin Alkin. Scriven's Modus Operandi Method and the Adversary Approach to evaluation are also discussed. The chapter on evaluation designs describes experimental designs, quasi-experimental designs, and process evaluation. Holistic evaluation and Transactional Evaluation are presented as integrated approaches to program evaluation. The second half of this monograph presents several case studies. They include evaluations of an equal educational opportunity program in the California Community Colleges, Project Head Start, a professional school curriculum, and public school curriculum; and needs assessments of a professional school and a faculty development program. The final chapter deals with the utilization of the results of an evaluation. A list of 112 bibliographical references is appended. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

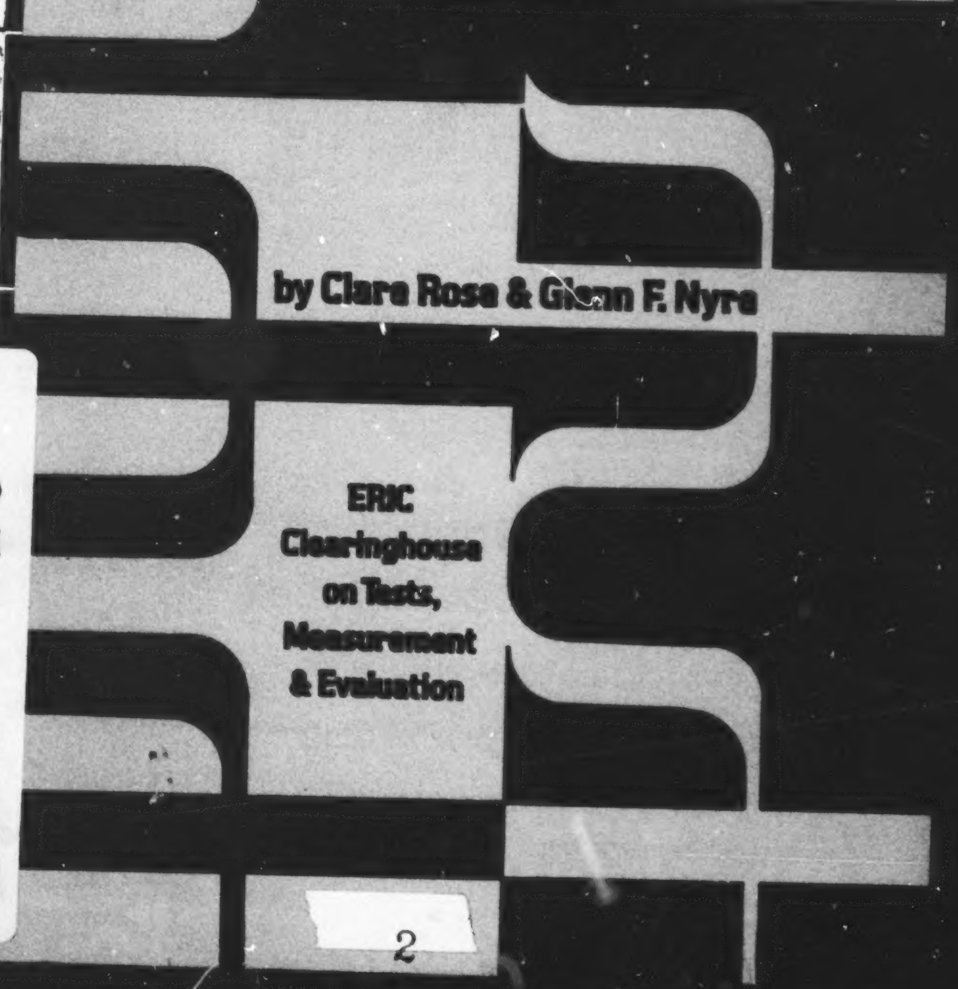
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED152859

TM007 218

THE PRACTICE OF EVALUATION

ERIC/TM Report 65



by Clare Rose & Glenn F. Nyre

**ERIC
Clearinghouse
on Tests,
Measurement
& Evaluation**

The Authors

Clare Rose is president and Glen Nyre is vice-president and executive director of the Evaluation and Training Institute, 11110 Ohio Avenue, Los Angeles, California 90025.

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either these reviewers or the National Institute of Education.

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service
Princeton, NJ 08541

December 1977

CONTENTS

PREFACE v

INTRODUCTION 1

EDUCATIONAL EVALUATION: ISSUES AND TERMS 2
  The Problem of Definition 5
  A Definition of Evaluation 7

MODELS OF EVALUATION 9
  The Countenance Model 10
  Goal Attainment Models 12
  The Discrepancy Model 14
  The CIPP Model 16
  The CSE Model 22
  Some New Approaches 24

EVALUATION DESIGNS 26
  Quasi-Experimental Designs 28
  Experimental Designs 29
  Process Evaluation—the Other Extreme 34

INTEGRATED APPROACHES TO PROGRAM EVALUATION 38
  Holistic Evaluation 38
  Transactional Evaluation 39

CASE STUDIES 42
  The Evaluation of Social-Action Programs 42
  EOPS: A Case Study of Holistic Evaluation 43
  Project Head Start: A Case of What Went Wrong 54
  Curriculum Evaluation 58
  An Evaluation of a Professional School Curriculum 60
  An Evaluation of a Public School Curriculum 67
  Needs Assessments 71
  Summary 74

UTILIZATION, QUALITY, AND ETHICS 76

REFERENCES 81

PREFACE

The literature of educational evaluation, consistent with its adolescence, seems to be smack in the middle of a growth spurt. The almost total paucity of such literature a decade ago has been supplanted by a goodly assortment of educational evaluation writings today. Unless, like the teenager it is, our evaluation literature suddenly stops growing, we can surely predict a geometric expansion of evaluation writings in the decades to come.

As usual, of course, tomorrow's evaluation literature will be markedly influenced by today's evaluation writers. Fortunately, Rose and Nyre have put together a monograph that should have a salutary influence on the literature to come. More immediately, it should prove useful to educators who are getting ready to wade into that real-world, cost-conscious, politicized, unpredictable maelstrom known as educational evaluation. It is a constant source of amusement to practicing educational evaluators that the uninitiated conceive of educational evaluation as largely an enterprise in which theoretical models are adroitly employed to cope with the realities of educational practice. After reading *The Practice of Evaluation*, it would be difficult to hold that view.

Rose and Nyre have divided their monograph into two essentially distinct segments, the first of which provides the reader with a succinct overview of the rudimentary theoretical concerns that educational evaluators have been tangling with for the past decade or so. For the beginner, this section will prove useful as an introduction to the field.

In the second, and to this reader the most interesting, section of the monograph, they describe a series of actual evaluations. These case studies are particularly intriguing because in all but two instances the authors are reporting on evaluations in which they personally took part. Few theoretical texts on evaluation can ever, with the candor employed here, capture so vividly the dilemmas faced by evaluators who are attempting to do an intellectually defensible job but must still tussle with the practicalities of life in the real world and all its pressures to compromise one's standards. Rose and Nyre offer us some useful insights into that world from the perspective of individuals operating a private evaluation agency.

The reader should become familiar with the theoretical discussions in the initial section of the book in order to make the subsequent case studies all the more meaningful. Interpreting real case studies according to theoretical propositions will, of course, make for difficult reading. But who ever said that educational evaluation ought to be easy?

W. James Popham
University of California, Los Angeles
and
Instructional Objectives Exchange

> *Any professional area that is so much avoided; that produces so many anxieties; that immobilizes the very people who want to avail themselves of it; that is incapable of operational definition, even by its most trained advocates, who in fact render bad advice to the practitioners who consult them; which is not effective in answering reasonable and important questions, and which has made little apparent effort to isolate and ameliorate its most serious problems—must indeed give us pause.*
>
> E. G. Guba

INTRODUCTION

Less than five years ago, our collection of non-journal works on evaluation consisted of a few well-worn monographs and even fewer books. Today, our file drawers and shelves are filled. There are well over a dozen hard-cover books complete with artist-designed jackets; most were written in the last two or three years. But, with all their instructional value, there is not one casebook among them that describes real-world evaluations in the context of recommended evaluation models and designs. After all the theory has been studied and the methodologies learned, only such a book can provide guidance to fledgling evaluators (or even seasoned ones) in the practice of program evaluation.

Although we, too, felt compelled to deal with basic principles, procedures, and methodological issues (and the first part of this monograph is devoted to their treatment), they are presented primarily as a foundation for the case studies that follow and simply provide increased understanding of why the evaluators carried out their investigations as they did. The purpose of this monograph is to provide an overview of basic principles and procedures and a guide to the practice of evaluation.

We first entered the world of evaluation with about equal proportions of good intentions, graduate training in research methodology, experience in survey research, high hopes, and naiveté. We were going to reform education through the wisdom and insight of our impeccably planned, exquisitely elegant evaluations.

The first evaluation we were asked to conduct involved a staff-training program for public school teacher specialists in an urban ghetto. The budget was miniscule, but we didn't care. When we asked about the purpose of the evaluation, we were told, "Every program should be evaluated." Here were our kind of people. They believed in the monumental and essential value of evaluation!

We developed (and even pretested) several forms of questionnaires and interview schedules. Because it was impossible to pin down any goals for the program, we were afraid we might overlook what could turn out to be valuable data. We spent long hours trying to figure out a way to cast the study into an experimental mode. But all of the teachers in the district office were going to participate (we had actually been called in *before* the program got under way, just as our professors had told us it should be done), and the idea of using a control group was ludicrous.

When we arrived at the site where the week-long program was to take place, we met the participants for the first time, saw the schedule of "activities," and held our breath. As it turned out, there was no staff development program and there never was any plan for an evaluation. We had become pawns in a political confrontation between two ethnic groups, who, in addition to warring against each other, had joined together to protest some of the district supervisor's policies. We had been hired as the final touch to distract the supervisor from the real purpose of the week—a showdown similar at least in emotion to the last walk in *High Noon*.

Certainly, this was a most unusual situation, but intent on our purpose, we had put blinders on to the tensions all around us. Fortunately, we have never encountered a similar case since then. But we have found ourselves in many situations where we could not randomize or identify comparison groups, and where input data were seldom available and school personnel resisted our pleas for performance testing to obtain outcome data. And we have been asked to conduct "formative" evaluations long after programs have been in operation.

Over the years, we have learned that for every program that permits rigorous and systematic data collection based on defined and generally agreed upon program goals, there are many more that are hotbeds of controversy with different groups of people holding different goals for the program and seeking different information from the evaluation. For every program that permits randomized assignment to treatment and control groups, there are many more in which the real participants of the program are hard to identify, let alone cast in an experimental design. And finally, we suspect that for every evaluator engaged from a project's inception in a well-planned, well-funded, potentially significant evaluation study, there are dozens more who find themselves faced with the task of evaluation in a far less ideal situation. These are the common problems encountered by people engaged in program evaluation. This monograph is addressed to them.

EDUCATIONAL EVALUATION: ISSUES AND TERMS

Evaluation is not a new concept; nor is it unique to education. Moses evaluated when he decided to risk the perils of foreign travel and led the people out of Egypt. David evaluated, albeit hurriedly, when he aimed the slingshot at Goliath's forehead. We all evaluate. Deciding whether to go to Europe or stay home and paint the house during summer vacation involves both affective and economic evaluation. When we go the the market to buy apples, we are evaluating as we select the largest, firmest, juiciest, and reddest (or greenest, depending upon your preference). Every time we make a decision, more or less rationally, systematically weighing the advantages and disadvantages of the alternatives, we are engaging in evaluation.

Formal evaluation has an equally long history, dating back to 2000 B.C. when Chinese officials administered civil service examinations (111). The first formal educational evaluation was conducted in the United States in 1887 by Joseph Mayer Rice, a free-thinking pediatrician. Considered a landmark study, in contrast to the simplistic surveys and even more simplistic interpretations that were characteristic of the time, Rice developed his own spelling test and administered it to over thirty thousand students in a large metropolitan school district. He wanted to show that student achievement had no relationship to the amount of time students spent in what he felt were senseless and interminable spelling drills (111). Unfortunately, a sophisticated technology did not evolve as a result of Rice's study, and most of the activity conducted in the name of evaluation for the next 20 or 30 years consisted of giving school children a variety of tests in every different subject. Measurement, not evaluation, leaped ahead.

It was not until the 1930s, when another trailblazer by the name of Ralph Tyler demonstrated a new approach to evaluation in the Eight-Year Study of the Progressive Education Association, that the foundation was laid for the form of evaluation we know today. Tyler conceived of evaluation as the process of determining the degree to which the goals of a program have been achieved. And, to Tyler, goals and objectives had to be defined in behavioral terms. Goals were derived from three basic sources: students, society, and the subject matter. General goal statements were then analyzed within the context of the psychology of learning (Can they be attained by the target population?) and a philosophy of education (Are they worthwhile and compatible with the purpose of education?). The goals that remain after this screening are transformed into specific behavioral statements of objectives; the degree to which students attain these objectives at the end of a program is measured; and the results are used to judge the effectiveness of the program (96). Goal-attainment models of program evaluation are much in evidence today and form the base of many experimental studies.

Still, the demand for formal program evaluation was not ignited until after the launching of the first Russian satellite. Sputnik will probably be remembered in the education world less for its impact on the space program than for its launching of the educational reform movement. Both educational reform and evaluation owe the beginnings of their modern histories to the furor created by the Russian feat. Public outrage turned against the schools, and for the first time in American history, the quality of our most honored institution, the school system, was seriously questioned. In part because of this concern, and in part because of civil rights groups' demands for fair treatment of minority children in the schools, the federal government began to contribute a greater share of the schools' financial support, which up until this time had been provided almost entirely by state and local governments. And, with the federal dollars came accountability. The federal government simply wanted to know if their money had been spent wisely. But the interest in accountability blossomed and culminated in the provisions for mandatory evaluation that were written into the Elementary and Secondary Education Act (ESEA) of 1965.

The ESEA, through its various titled programs, provided for thousands of grants to educational agencies throughout the country, and each local project had to be evaluated in order to continue receiving federal funds. Not surprisingly, the educational community was not equipped to handle the vast numbers of evaluations that were required to satisfy the law. Professional evaluators did not yet exist, and few educators were knowledgeable about evaluation. Academics trained in research or measurement were drafted to conduct the evaluations, and they approached the task as researchers, not as evaluators. Masses of unnecessary data filled the volumes of project reports, and, not surprisingly, the federal government found them to be of little help.

Large-scale evaluations of federal programs fared no better. Would-be evaluators clung tenaciously to the classical experimental model with which they were familiar. Strict adherence was given to defining program goals, usually in tandem with a list of null hypotheses; assigning subjects randomly to experimental and control groups; collecting masses of data from each group, usually in the form of standardized achievement measures; employing statistical techniques of varying degrees of sophistication; and, finally, making judgments regarding the worth of the program based on a comparison of the two groups. Comparisons of randomly assigned treatment and control groups became the *sine qua non* of program evaluation. Unfortunately, the emphasis on testing and the collection of quantitative data caused many people to confuse measurement, accompanied by vast amounts of "illustrative" data, with evaluation—a confusion that continues to exist even today.

The deficiencies of experimental design are discussed in detail in a later section, but it is sufficient to say at this point that the evaluation reports they provided were dismal failures. Used by graduate schools today as examples of what *not* to do in program evaluation, these comparative studies yielded "no statistical differences" over and over again. Program budgets were cut or eliminated out of political expediency alone; others continued business as usual without a shred of evidence as to their effectiveness.

The shadows cast over evaluation as a result of these early studies have remained and in many ways have influenced recent trends in evaluation practices. Nevertheless, the Elementary and Secondary Education Act of 1965 must be credited with providing the impetus for evaluation, an activity that has turned out to have had an equal, if not greater, impact on education than the act itself.

The Problem of Definition

From these inauspicious beginnings emerged the field of evaluation as we know it today—a field that is characterized by confusion, conflict, controversy, and mistrust. Evaluators do not share a common philosophy, focus or terminology. Fiercely loyal to different "schools" of evaluation, educators argue over *goal-free*, *goal-based*, and *formative* and *summative* evaluation. Even the most basic terms, such as *measurement*, *assessment*, and *evaluation* are used interchangeably and often incorrectly. It is no wonder that in some quarters evaluation is not yet legitimized. In order to clarify some of the major evaluation terms with which the reader should be conversant, it will be helpful to examine their definitions before we proceed with our discussion.

Accountability: Accountability is concerned with furthering the educational effectiveness of school systems (3). The *Random House Dictionary of the English Language* shows the synonym of accountability to be "responsibility." Educational accountability thus represents the educators' acceptance of responsibility for the consequences of the educational system entrusted to them by the public. Evaluation is an intrinsic part of accountability. Program effectiveness must be evaluated to provide information for teachers, administrators and program directors, as well as legislators and other officials who allocate the funds for the programs and for the public who provides the funds through their tax dollars. Accountability is usually a condition *requiring* evaluation; but accountability is not equivalent to evaluation.

Measurement: As we said earlier, measurement is often equated with evaluation, since so many of the early evaluation reports consisted primarily of measurement data. But measurement is static—it is the act or process of determining the extent, dimensions, quantity, or capacity of something at one point in time. In education, measurement is the act of determining the extent to which an individual has learned or the degree to which an individual possesses a certain characteristic, ability, or talent. Measurement is usually part of the evaluation process, providing useful data for evaluation, but again, the two terms are not equivalent.

Assessment: Like measurement, the term assessment is often used interchangeably with evaluation, and several major evaluation projects have been referred to as "National Assessments." Assessment is really more akin to measurement, however, and refers to the process of gathering and collating the data. Anderson and associates (3) claim that assessment has a narrower meaning than evaluation and a broader meaning than measurement. In addition to the act of measurement, assessment involves the qualitative judgment of determining what and how to measure as well as the process of putting the data into an interpretable form.

Evaluation Research: Although many writers classify evaluation as a form of research, or conversely, view evaluation research as a specific method of evaluation, others make a sharp distinction between the two terms. Evaluation research is defined as the application of social science methods to discover information of importance to program practice and public policy (98). Implicit in the distinction is that the evaluator doing evaluative research acts as an objective scientist, employing quantitative and reproducible techniques and eschewing judgment. Research is primarily concerned with the basic theory and design of a program over a set period of time. Evaluation may to some extent be concerned with basic theory and design, but its primary function is to appraise a program to determine its merit.

Formative and Summative Evaluation: Coined by Michael Scriven[1] (76), these terms distinguish between the two basically different roles served by evaluation. Formative evaluation refers to those evaluations undertaken *during* the developmental process for the express purpose of guiding and assisting program improvement. In a formative evaluation, the evaluator might gather specific data on various aspects or components of the program at several stages throughout the developmental phase in order to identify areas requiring improvement. This information provides the developer with empirical data to help determine where and how to revise the program and make it better.

Summative evaluation, on the other hand, refers to the final evaluation of a program and is concerned with determining the worth of the overall program *after* it has been completed. The purpose of summative evaluation is to help make decisions regarding the program's future—its continuance, termination, replication and/or dissemination. Implicit within these two terms, formative and summative, is another distinction, which refers to the evaluator's role. That is, because the purpose of formative evaluation is *to improve*, the formative evaluator becomes part of the developmental process and the task of formative evaluation can even be performed by the program developer. If a person other than the developer performs the work of formative evaluation, that person can work closely and collaboratively with the developer. The point is that there is no need to ensure third-party objectivity in the formative stages of program development. The goal is improvement, and both the developer and evaluator can be committed to that end. The summative evaluator is in a different position. Summative or final, end-of-program evaluation demands an objective and impartial evaluation, since the future of the program is at stake. The summative evaluator must be completely independent of the developer.

---

[1] Formative evaluation, as described by Scriven, is similar to what Cronbach (19) talks about in his discussion of evaluation for course improvement, although he and Scriven strongly disagree as to the relative importance of the role of formative evaluation, with Cronbach taking the position that formative is of greater importance than summative evaluation.

Although these terms were developed for the evaluation of curriculum materials, they have been adopted by the educational community as part of the basic vocabulary of evaluation and are used to distinguish the two operations in any type of evaluation enterprise.

Goal-free Evaluation: Another term created by Scriven (77), goal-free evaluation is an approach that aims to ensure that evaluators pay attention to the actual outcomes of a program, intended as well as unanticipated, rather than just the quality of the program goals or the extent to which they have been achieved. Scriven was concerned that an evaluator would become preoccupied with goals and, consciously or unconsciously, ignore the wide range of actual outcomes which, intended or not, are nevertheless real. In the goal-free approach, the evaluator deliberately avoids gaining any knowledge of the program goals (a simple task in cases where program goals don't really exist), gathers data on the actual outcomes only, and then evaluates their importance. Goal-free evaluation was not conceptualized to replace goal-based evaluation, but to augment it and thus provide a more reliable and valid evaluation.

Goal-based Evaluation: Goal-based evaluations refer to evaluations that are based on the extent to which intended project goals have been achieved. As suggested by Scriven, this should be accompanied by an assessment of the quality of the goals established in the first place (76).

A Definition of Evaluation

Finally, the most important term to define, and one of the most controversial, is the word evaluation itself. The attempt to clarify the meaning of evaluation is not an idle exercise. Quite the contrary. It is of major importance since no one is agreed upon a definition and the different definitions people accept carry with them different advantages and disadvantages, each affecting the way in which evaluators approach and carry out their tasks. For example, three definitions of evaluation have appeared at one time or another in its history: measurement; congruence between objectives and performance; and judgment (59). When measurement is accepted as the definition of evaluation, the evaluator's main task is to administer tests and gather measurements. The role of the evaluator is equivalent to that of psycnometrist. If evaluation is defined as professional judgment, then a group of "experts" would observe a program in action and, subsequently, pronounce judgment expertly—an act reminiscent of accreditation procedures from whence the definition is derived.

Definitions of evaluation also provide the conceptual base for the models of evaluation, and, although there are still a few educators who subscribe to the measurement definition (23, 93), an examination of the literature and a review of the different models and classification schemes indicate that model builders and evaluation writers cluster around three major definitions: 1) those that define evaluation as an assessment of the discrepancy between objectives and performance (Metfessel and Michael; Provus; Stake; Tyler); 2) those that focus on outcomes and define evaluation as an assessment of outcomes, intended or otherwise (Popham; Scriven); and 3) those who are decision oriented, defining evaluation as the process of obtaining and providing information for decision makers (Alkin; Cronbach; Guba and Stufflebeam). Each of these "schools" of evaluation thought and the writings of their proponents will be discussed subsequently.

A central issue for all three groups is that of value. The advocates of judgment follow the dictionary definition, which states that "to evaluate is to ascertain the value of" (*Random House Dictionary*). Thus, Popham (60) speaks of formal evaluation as the "assessment of the worth of educational phenomena" and Scriven (76) goes further, suggesting that without judgment of merit, no evaluation has taken place. Glass similarly stresses that evaluation is an attempt to assess the worth or social utility of a thing, and Stake (83) specifies description and judgment as the two basic ingredients of evaluation. Dressel (21) broadens the definition to include process. To Dressel, evaluation is "both a judgment on the worth or impact of a program, procedure or individual and the process whereby that judgment is made." Others who support the judgment of merit position include Airasian (1), Sax (73), Suchman (91), Weiss (98, 99), and Wholey *et al.* (107).

At the other end of the spectrum are those who eschew a value orientation, viewing the function of evaluation instead within the context of decision making only. In this case, the evaluator gathers information concerning the relative advantages and disadvantages of various decision alternatives so that decisions can be made rationally and systematically. The uses to which evaluation information is actually put by decision makers is yet another matter, one that will be dealt with later. Guba and Stufflebeam (37) object to judgment or value definitions because they ignore the processes of arriving at the information. They suggest instead that "evaluation is the process of delineating, obtaining and providing useful information for judging decision alternatives." Along the same lines, Alkin (2) offers a somewhat longer and broader version, which includes identifying the decision areas as well as collecting and providing the information to decision makers.

Some who oppose the value dimensions are concerned that passing judgment will ultimately diminish the evaluator's access to data and evaluation will become even more suspect than it is now. Others, such as Guba and Stufflebeam, Provus, and Alkin, take the position that the act of judging or making the final determination of the worth or merit of an educational program or product is only within the purview of the decision maker, not the evaluator. Popham (60) refers to the three models upon which these definitions are based as "decision-facilitation models." Although they do involve the evaluator's use of judgment as well as a determination of whether the program goals have been attained, their orientation is toward servicing decision makers. "The orientation of these models is so overwhelmingly toward servicing educational decision-makers that some of their proponents conceive of the evaluator as the decision-maker's handmaiden/handmister." (60) Brief descriptions of these models are presented in the next section.

MODELS OF EVALUATION

Evaluation models are as prolific as rabbits, and they procreate about as speedily. No longer do people develop an idea or test an approach. Instead, they develop a model. Often spawned from combinations of several other models, some from other disciplines, they become progressively more grandiose in their complexity, more esoteric in their terminology and more pompous in their names. One has only to examine a recent program schedule for the American Educational Research Association's (AERA) annual meeting or the extensive Educational Resources Information Center (ERIC) abstracts on evaluation. The most frequently used paper title begins with the words "The Development of an Evaluation Model for . . . ."

The array of evaluation models from which we may choose would, if nothing else, provide a marvelous tongue-twisting party game. Just imagine what it would sound like if someone who'd had too much to drink were to chant in mantra form the names of evaluation models and approaches. We have democratic evaluation, responsive evaluation, transactional evaluation, modus-operandi evaluation, holistic evaluation, discrepancy evaluation, goal-free evaluation, and adversary evaluation. There is the Countenance Model, the Differential Evaluation Model, the Priority Decision Model, the Trade-Off and Comparative Cost Model, the Systems Approach Model, and the Cost Utility Model. There are Ontological Models, Synergistic Models, and Ethnographic Models.[2] And this is only a partial list. Indeed, model building has become so commonplace, that to be truly distinctive these days one should eschew model molding altogether.

Many of these so-called models, of course, are not really models, but rather, descriptions of processes or approaches to program evaluation. The purpose of a model is to guide and focus inquiry. Borich (7) indicates that models in the social sciences have three identifiable characteristics: precision, specificity, and verifiability. Models are precise because they are quantitative in nature. The elaborate forms of measurement are derived purposefully to describe the phenomena under investigation. Models are specific because they deal with only a certain number of phenomena. Models are verifiable in the sense that hypotheses are formulated and empirical evidence is accumulated that eventually determines the model's accuracy and usefulness. In listing the criteria for models, Carter (13) suggests that they must be efficient, heuristic, internally logical and complete; capable of being extended by empirical study; capable of helping the evaluator anticipate all of the information needs for decision making and capable of relating elements in ways not previously related. Borich (7) hastens to add that while "evaluators strive to construct models that are precise, specific and verifiable, the end result often falls short of that which can be expected in the sciences." Models are, in effect, conceptualizations, and they may be theoretically sound; but they do not necessarily lend themselves to actual implementation.

---

[2] The models and their authors are listed at the end of this section to avoid interrupting the flow of the text.

A few models were no doubt built by Rube Goldberg fans intrigued by mazes of convoluted lines, arrows, and dots, and even the best of models are not perfect. Still, this should not deter would-be evaluators from having in their repertoire an understanding of the major evaluation models that have been dominant in the literature and influential in the field. We will examine a few of the important models that have guided evaluations during the last few years.[3]

The Countenance Model

Created by Robert Stake (85), the Countenance Model is so named because of the title of his article describing it ("The Countenance of Educational Evaluation"). This model is based on the notion that judgment and description are both essential to the evaluation of educationl programs. Accordingly, Stake distinguishes between three bodies of information that are elements of evaluation statements that should be included in both descriptive and judgmental acts. These elements are: antecedents, transactions, and outcomes.

Antecedents refer to conditions existing prior to implementation of the program that may relate to outcomes. Transactions are the "succession of engagements" that constitute the process (in other words, the instructional process or educational aspect of the program). Films, examinations, homework, class discussions, and teachers' comments on student papers are all examples of transactions. Outcomes, as conceived by Stake, refer to much more than traditional student outcomes. They include immediate, long-range, cognitive, affective, person, and societal outcomes. Outcomes also include the program's impact on teachers, administrators, and others as well as the wear and tear on equipment and facilities in its conduct.

---

[3] For comparative analyses of the different models, readers are referred to Worthen and Sanders' (111) multi-page descriptive matrix of models; Wetherill and Buttram's (105) comparison of 21 models, and Carter's (13) taxonomy of decision-oriented evaluation models.

Descriptive information is classified either as *intents* or *observations*. Intents include program objectives—not only intended student outcomes, but also the planned-for environmental conditions as well. The judgment matrix includes both the standards used to reach judgments and the actual judgments themselves. A graphic representation of Stake's layout is presented in Figure 1.

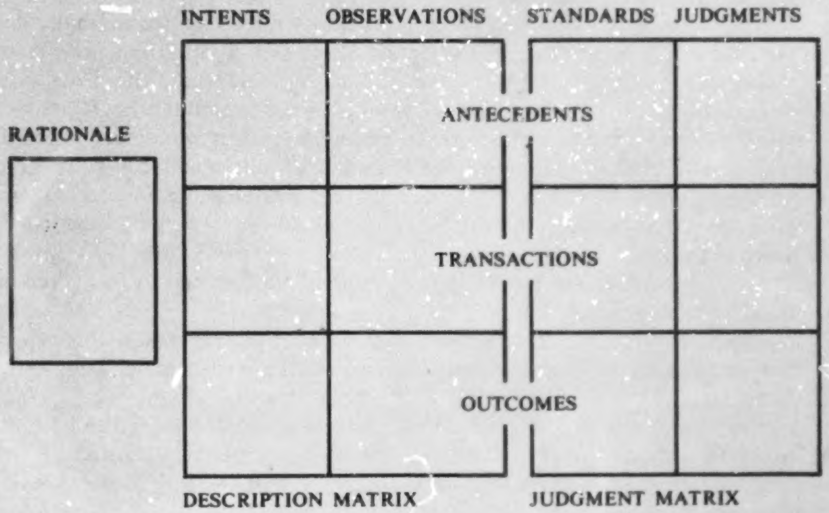


Figure 1. Layout of the Countenance Model*

Note that a separate box depicted to the left of the layout is labeled *rationale*. According to Stake, an evaluation is not complete without a statement of the program's rationale. This statement indicates the philosophical background and basic purposes of the program and provides a basis for evaluating intents.

There are two principal ways of processing descriptive evaluative data: finding the contingencies among antecedents, transactions, and outcomes; and finding the congruencies between intents and observations. The data for a program are congruent if what was intended actually happened, although Stake admits that it is unlikely that all of the intended antecedents, transactions, and outcomes come to pass exactly as intended even in the best of programs. With reference to transaction data, Stake insists that the evaluator carefully observe and record data emerging from the transactional and interactional classroom processes. He broadens the general concept of out-

---

*W. James Popham, *Educational Evaluation*, © 1975, p. 31. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

come data to include future application, transfer, and the effect of process on outcomes.

The contingencies among the variables are of special importance to the evaluator. In the sense that evaluation is the search for relationships that facilitate educational improvement, the countenance evaluator's task is to identify outcomes that are contingent upon particular antecedent conditions and instructional transactions.

We previously stated that the foundation for a model's orientation derives from the author's definition of evaluation. In this case, Stake is a proponent of the value-judgment school; the model is judgmental and the process of judging the merit of a program is an integral part of the model. There are two bases for judging the characteristics of a program in the Countenance Model: evaluating a program either on the basis of absolute standards or relative standards—that is, either standards reflecting personal opinion concerning what the program should be or standards reflecting other similar programs. Judgment is involved in choosing which set of standards to use—absolute or relative—to obtain an overall rating of merit upon which to base recommendations regarding the future of the program.

In later writings on "responsive evaluation," Stake (84) adds that rather than personally passing judgment, the evaluator should collect samples of the judgments of many people in the program—the clients, staff, community, and others.[4] Stake's emphasis on the evaluator's need to be fully aware of and sensitive to the concerns of many people affected by the program became the central theme in several "process-only" evaluation approaches discussed in the next chapter.

Goal Attainment Models

Fathered by Ralph Tyler in the 1930s, goal-attainment or objectives-oriented models still provide guidance for many evaluations and occupy an important place in the literature. An example of a goal-attainment model is the paradigm developed by Metfessel and Michael (54). The steps of their model are:

1. Involve members of the total community directly and indirectly as participants in the evaluation;
2. Develop broad goals and specific operational objectives, both cognitive and noncognitive;
3. Translate objectives into forms that are communicable and that can be implemented to facilitate learning;

---

[4] Many prominent evaluation theorists expanded the classic paradigm by broadening the definition of decision maker and legitimizing data other than test scores, particularly the judgments of various people involved directly and indirectly with the program (75).

4. Develop criterion measures and instruments to determine whether the program achieved the objectives;
5. Measure the program's progress toward attainmnent of the objectives and, finally, measure attainment of the objectives;
6. Analyze the data;
7. Interpret the data in light of established standards and values; and
8. Formulate recommendations for program improvement as well as for revisions in the goals and objectives.

The appendices to the article contain lists of criterion measures (for which Metfessel and Michael have become better known than for their paradigm) that can be used by the evaluator in the fourth step of the model. The measures are wide-ranging, with those for determining student behavior including self-inventories, standardized tests, rating scales, projective tests, anecdotal records and case histories. Measures are also provided for teacher and community behavior.

Somewhat similar to Metfessel and Michael's strategy is one offered by Robert Glaser (29). His scheme, which excludes summative evaluation, consists of six steps that comprise a continuing cycle of formative evaluation:

1. Specify the outcomes of learning in measurable terms;
2. Analyze the learners' entry behavior—the level of knowledge, skill, or ability already in the students' repertoire relevant to each task specified in the objectives;
3. Provide students with various learning alternatives;
4. Monitor students' progress toward objectives;
5. Adjust the instructional program according to the level of students' performance as they progress toward attainment of the objectives; and
6. Evaluate the program for on-going feedback and program improvement.

Glaser's paradigm is most suited to the evaluation of instructional programs, although the strategy is generalizable to other program situations. Glaser has been particularly effective in specifying the conditions necessary for the evaluation of instruction, and his main contribution in this area is his emphasis on detailed diagnosis of student (participant) entry behaviors, an emphasis that is important in almost all program evaluations.

Despite their several advantages, there are more than a few criticisms of goal-attainment models. Scriven (76) was the first to caution against indiscriminate goal-based evaluation without an accompanying evaluation of the quality of the goals themselves: ". . . it is obvious that if the goals aren't worth achieving then it is uninteresting how well they are achieved." Unfortunately, many evaluators do not heed Scriven's advice, and the goals established for a program often remain unscrutinized.

Another major problem with goal-based models is that in order to provide an effective base for determining program results, program objectives must be clear and specific. Rarely are evaluators afforded the luxury of explicit program goals. More often than not, if they exist at all, the objectives are vague, general, and too broad to provide a base for comparing results. Dressel (21) offers a reasonable explanation for the prevalence of globally stated program objectives, simply stating that "it is far easier to generate agreement among different constituent groups if an objective is vague." Broad goals are seldom controversial. For example, few people would argue if the goal of a program were to enhance students' self-confidence or improve their ability to relate to people or other such incontrovertibly inspiring goals. Agreement concerning the behaviors or attitudes that students would have to demonstrate in order to show that they had indeed increased their self-confidence or their ability to relate to people would be far more difficult to obtain. In fact, whether or not objectives of this type can even be defined in specific measurable terms is itself a subject of great controversy.

A third, frequently heard, criticism of goal-based evaluations is that focusing attention on the results of a program only in terms of its intended objectives narrows the evaluation, so that the different procedures used to achieve the results and their relationship to program outcomes are ignored. Global judgments of merit, of course, can be made concerning the overall value of the program as far as its success in achieving the objectives is concerned, but no basis for program improvement—an equally important part of evaluation—can be provided by the data. In other words, the goal-attainment model is not decision oriented; only limited information can be provided for decision makers. In decision-oriented models, the purpose of evaluation is to provide information for decision makers for a multiplicity of decisions—decisions concerning whether or not a program is needed in the first place; decisions about whether to continue, expand, or terminate a program; decisions concerning program certification or licensing; and decisions about program improvement. The next two models that are described qualify as decision-oriented models for program evaluation, an orientation that is evident in the definition of evaluation that provides the conceptual base for their development.

The Discrepancy Model

A very popular and widely used model is Malcolm Provus' Discrepancy Model, so named because the discrepancy between performance and standards is a key point in his definition of evaluation. Provus (64) defines evaluation as:

> . . . the process of 1) defining program standards; 2) determining whether a discrepancy exists between some aspect of program performance and the standards governing that aspect of the program; and 3) using discrepancy information either to change performance or to change program standards.

Depending upon the information yielded as a result of the evaluation, there are four possible decisions to be made. The program can be terminated; it can be modified; it can continue or be repeated as is; or the standards can be changed.

The Discrepancy Model involves five stages, each of which involves a comparison between reality, or performance, and standards. Discrepancies are determined by examining the three content categories (input, process, and output) at each stage and comparing the program performance information with these defined standards at each stage.

The design of the program is compared with design criteria; program operations are compared against the input and process sections of the program design; the degree to which interim objectives are achieved is compared with the relationship between process and product; the achievement of terminal objectives is compared with their specification in the program design; and, finally, the cost of the program is compared against the cost of other programs with similar goals.

The first stage focuses on the *design* and refers to the nature of the program—its objectives, students, staff and other resources required for the program, and the actual activities designed to promote attainment of the objectives. The program design that emerges becomes the standard against which the program is compared in the next stage.

The second stage, *installation,* involves determining whether an implemented program is congruent with its implementation plan. *Process* is the third stage, in which the evaluator serves in a formative role, comparing performance with standards and focusing on the extent to which the interim or enabling objectives have been achieved. The fourth stage, *product,* is concerned with comparing actual attainments against the standards (objectives) derived during Stage 1 and noting the discrepancies. The fifth and final stage is concerned with the question of *cost.* A cost-benefit analysis is made of the completed program and compared to other programs similar in nature.

Because the primary function and orientation of the Discrepancy Model is to provide information for decision makers, Popham classifies it in his four-part model medley as a "decision-facilitation" model (60). But, as Popham acknowledges, there is overlap between the categories, and the Discrepancy Model is vulnerable to the same criticisms leveled at the goal-attainment models.

The CIPP Model

One of the most well-known and widely used models is the CIPP Model developed by Egon Guba and Daniel Stufflebeam (37). CIPP is an acronym that stands for the four types of evaluations for which the model is appropriate: *context* evaluation, *input* evaluation, *process* evaluation, and *product* evaluation.

As noted earlier, the foundation for the development of a model is the author's definition of evaluation, and for Guba and Stufflebeam "evaluation is the process of delineating, obtaining and providing useful information for judging decision alternatives."

This definition contains three important points. First, evaluation is a systematic, continuing process. Secondly, the process includes three basic steps: 1) delineating the questions to be answered; 2) obtaining relevant information so that the questions may be answered; and 3) providing the information for decision makers. Thirdly, evaluation serves decision making. Although there is a judgmental component, the primary emphasis in this model is on decision making. Basically, the CIPP model answers four questions: 1) What objectives should be accomplished? 2) What procedures should be followed in order to accomplish the objectives? 3) Are the procedures working properly? and 4) Are the objectives being achieved?

The CIPP Model, pictured in Figure 2, distinguishes between four different decision-making settings in education and four corresponding types of decisions, in addition to the four types of evaluation that form the model's name. The first distinction, that of decision-making settings, arises directly as a consequence of the authors' definition of evaluation; that is, the extensiveness of an evaluation, as well as the rigor with which it is conducted, are determined in large measure by the importance of the decision that is to be serviced. The importance of the decision, in turn, depends upon the significance of the change it is intended to bring about. For example, decisions that will have far-reaching consequences demand evaluations that are thorough, rigorous, and, most likely, expensive. Decisions that will have little impact on the people or the system, such as the decision to change the entrance of a building, do not require expensive, detailed evaluations.

A second factor to be considered is the availability of information and the decision maker's ability to use it. Evaluations must, of necessity, be more extensive when there is little information already available or when the decision maker is not able to make use of the available information in its present form. These two factors—significance of the intended change and the availability of information, as well as the decision maker's ability to use it—form two intersecting lines which, when combined, yield four classes of decision settings. The continua are labeled "small versus large change" and "high versus low understanding." The rule for distinguishing between small and large change is the degree of controversy over the change. The more con-

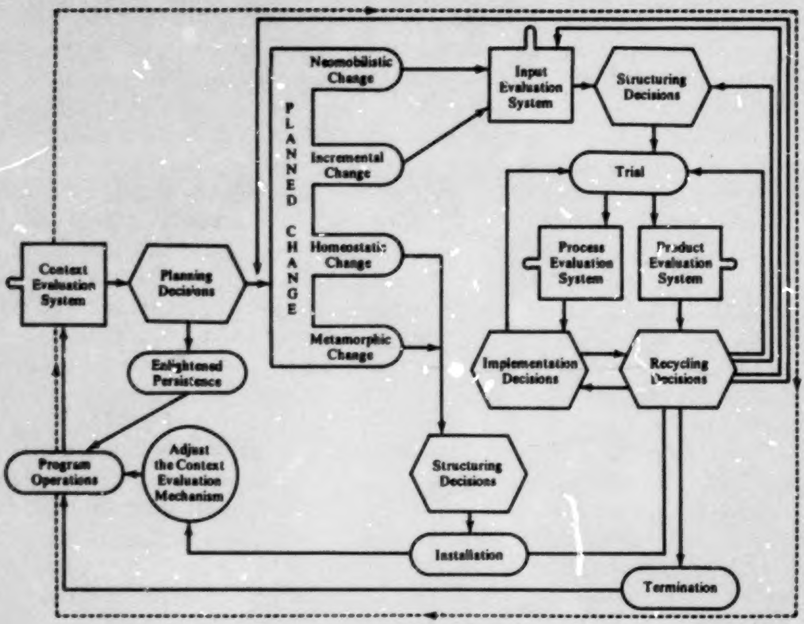

Neomobilistic Change
Input Evaluation System
Structuring Decisions
P L A N N E D C H A N G E
Incremental Change
Trial
Homeostatic Change
Process Evaluation System
Product Evaluation System
Context Evaluation System
Planning Decisions
Metamorphic Change
Implementation Decisions
Recycling Decisions
Enlightened Persistence
Program Operations
Adjust the Context Evaluation Mechanism
Structuring Decisions
Installation
Termination

Figure 2. The CIPP Model*

troversial the change, the larger or more important it is. School integration is a good example of a large, controversial change. Large changes usually involve major restructuring within the educational system.

Small changes, conversely, refer to changes that have no significant impact on variables considered to be important by society. Thus, small changes are relatively inconsequential and noncontroversial. Changing textbooks, however, or adding curricular content are examples of small changes that still require evaluative information for sound decisions.

The four decision settings are called homeostatic, incremental, neomobilistic, and metamorphic, each referring to the extent of intended change. Homeostatic decisions are aimed at maintaining the *status quo* and, not surprisingly, are characteristic of most decisions that are made in education. Faculty assignments and course scheduling are examples of homeostatic decisions. Incremental decisions refer to developmental activities, particularly those conducted as a part of continuous program improvement. Contrary to their creators' view, many innovations in education are examples of incremental activities—attempts to make some improvement without risking a major upheaval.

---

*Source: Phi Delta Kappa, National Study Committee on Evaluation. *Educational evaluation and decision making*. Ithaca, Ill.: Peacock Press, 1971. Reprinted by permission of Phi Delta Kappa, Incorporated.

Neomobilistic decisions denote large, innovative activities conducted for the purpose of solving significant problems. Policy research centers and institutes that deal with long-range educational planning are engaging in the area of neomobilistic decision-making. Metamorphic decision-making aims to produce complete changes in an educational system. Ivan Illich's proposal to disestablish schools is a good example of what would be metamorphic change in education. Quite obviously, this kind of change would be utopian, and the probability of its taking place in education is indeed slim.

Within each of these decision-making settings, there are thousands of specific educational decisions that are categorized by the authors into another foursome: 1) planning decisions to determine objectives; 2) structuring decisions to design the means or procedures to be used to attain the objectives; 3) implementing decisions to watch over and refine the procedures; and 4) recycling decisions to judge and react to the outcomes or attainments of the objectives.

Corresponding to each of these four decision types are the four types of evaluation for which the model was named—context, input, process, and product. Context evaluation is the most prevalent type of evaluation used in education. The major objective of context evaluation is to determine needs, specify the population and sample of individuals to be served, and devise objectives designed to meet these needs. The procedures for context evaluation include: 1) defining and describing the environment in which the change is to occur; 2) identifying unmet needs and necessary and available resources; 3) identifying sources of problems or deficiencies in meeting these needs; and 4) predicting future deficiencies by considering the desirable, expected, possible, and probable outcomes. In other words, context evaluation provides the rationale for justifying a particular type of program.

Context evaluation, according to Stufflebeam (90), addresses these questions:

1. What unmet needs exist in the context served by a particular institution?
2. What objectives should be pursued in order to meet these needs?
3. What objectives will receive support from the community?
4. Which set of objectives is most feasible to achieve?

Unmet needs can be determined by examining the goals of the school and students' performance, comparing them, and noting any "discrepancies." The differences represent unmet needs. Which objectives should be pursued in order to meet these needs depends on the conditions that account for the differences. Stufflebeam suggests that literature published by other evaluators who have experienced similar problems may help to explain why students failed to reach desired criterion levels. Which objectives will be supported by the community can be determined simply by polling or interviewing representatives of community groups. Determining which objectives are most feasible involves estimates of costs and of resources available to the school and community.

The purpose of input evaluation is to determine how to use the resources in order to meet the goals established for the program. The end product of input evaluation is an analysis of alternative procedural designs or strategies in terms of their potential costs and benefits.

Stufflebeam (90) suggests five questions that input evaluation should be capable of answering:

1. Does a given project strategy provide a logical response to a set of specified objectives?
2. Is a given strategy legal?
3. What strategies already exist with potential relevance for meeting previously established objectives?
4. What specific procedures and time schedules will be needed to implement a given strategy?
5. What are the operating characteristics and effects of competing strategies under pilot conditions?

Decisions based upon information collected in input evaluations typically result in the specification of materials, procedures, time schedules, facilities, staffing, and budgets that will be necessary to promote attainment of a particular set of objectives.

Process evaluation provides continuing, periodic feedback to program managers on how the project is progressing once it has been initiated. The objective of process evaluation is to detect defects in the design or its implementation and to monitor the various aspects of the project so that potential problems or sources of failure can be identified and remedied. As in formative evaluation, the process evaluator collects information frequently and reports it to the program manager as often as necessary to keep the project progressing as planned.

Stufflebeam (90) suggests the following questions to be addressed by process evaluation:

1. Is the project on schedule?
2. Should the staff be retrained or reoriented prior to completion of the present project cycle?
3. Are the facilities and materials being used adequately and appropriately?

4. What major procedural barriers need to be overcome during the present cycle?

In addition to providing feedback for ongoing program improvement, process evaluation yields a record or diary of the project which itself can prove valuable once the project has been completed.

Finally, product (or outcome) evaluation measures and interprets attainments at the end of a program and at appropriate cut-off points within it. Product evaluation includes: 1) identifying congruencies and discrepancies between the intended objectives and actual attainments; 2) identifying unintended results, desirable or otherwise; 3) providing for objectives that have not been met by recycling the program; and 4) providing information for decision makers regarding the future of the program—whether it should be continued, terminated, modified, or refocused.

Despite the labyrinthian intricacy of the model and the perhaps needlessly complex terminology, the CIPP model has been used extensively to guide program evaluations throughout the field of education (18, 28, 39). It was one of the first full-scale models that directed attention to the information needs of decision makers. The CIPP model made evaluators aware of both the variety and range of evaluative information that is necessarily a part of the different types of decisions that have to be made in education and the different settings in which those decisions have to be made.

In later works, Stufflebeam (88, 89) distinguished between evaluation for decision making and evaluation for accountability. Evaluation conducted for the purpose of decision making is proactive—similar in concept and practice to formative evaluation. Evaluation for the purpose of accountability is retroactive in nature and serves a summative role. Actually, all four types of evaluations—context, input, process, and product—can be considered formative when they provide information for program improvement and summative when they provide information for decisions regarding a program's future.

Dressell (21) illustrates this quartet within the context of the four corresponding parts of an educational program—input, environment, process, and output. Context evaluation contributes to decisions regarding the environment, but it is also concerned with the interrelations of all of the program parts. Input evaluation is concerned with clarifying goals and assessing the use of resources. Process evaluation corresponds to the process elements, analyzed in terms of their contribution to the attainment of objectives. Output evaluation determines the discrepancy between intent and reality and analyzes the factors contributing to the differences.

Although Guba and Stufflebeam do not provide a set of designs to accompany the four types of evaluation their model accommodates, they do offer a checklist of procedures for developing a design applicable to any of the four types. The checklist consists of six major steps: 1) focusing the evaluation, which means identifying and defining the decision situations or the goals of the evaluation, the setting within which it is to be conducted, and the policies within which it is to operate; 2) planning the data collection; 3) planning the organization of the data; 4) planning the data analysis; 5) specifying audiences, formats, means, and schedules for reporting the findings; and 6) administrating the evaluation, or providing an overall plan for executing the evaluation design. Dressel (21) offers a more comprehensive and useful checklist for planning an evaluation.*

A. What is the purpose and background of the evaluation?

   1. What inputs, environmental factors, processes, or outcomes are to be evaluated?
   2. What are the critical points at which evidence will be required for decisions?
   3. What rules, procedures, assumptions, and principles are involved in the decisions?
   4. Who will make decisions and what is the process by which these will be made?
   5. Does the overall situation suggest, require, or prohibit certain tactics and strategies?
   6. What timing considerations are involved?
   7. What are the limitations on costs?
   8. What are the specific evaluation tasks?

B. What information is to be collected?

   1. Are the particular items unambiguously defined and collectible by objective and reliable means?
   2. From where or from whom is the evidence to be collected?
   3. By whom is it to be collected?
   4. What instruments or procedures are to be used?
   5. Will the collection of evidence in itself seriously affect the input, environment, process, or outcomes?
   6. Will the collection of evidence become a regular part of the process, or is it an add-on for a one-time evaluation?
   7. What is the schedule for collection of information?

---

*Paul L. Dressel, *Handbook of Academic Evaluation*, © 1976, pp. 23–25. Reprinted by permission of Jossey-Bass, Inc., San Francisco, Calif.

C. What procedures will be used for organizing and analyzing data?

  1. In what form is information to be collected?
  2. Will coding be required? If subjective judgments will be required in coding, are the criteria for these adequate? Who will do the coding?
  3. How will the data be stored, retrieved, and processed?
  4. What analytic procedures are to be used?

D. Is the reporting procedure clear?

  1. Who will receive reports?
  2. Will reports be organized by analytic procedures, by type of data, or by decisions to be made?
  3. Will reports include the practical implications regarding the various possible decisions to be made or leave these implications for the project staff or administrators to ascertain?
  4. Is the evaluator to state explicitly the particular decisions which he believes are supported by the evidence?
  5. When and in what detail are reports to be made?

E. How is the evaluation to be evaluated?

  1. Who will be involved—project staff, the evaluator, decision-makers, some presumably more objective individual?
  2. What will the criteria used in this second-level evaluation be—costs, program improvement, impact on further planning of related enterprises?
  3. To whom and when is this report to be presented?
  4. What decisions are to be anticipated as a result of the report? Will they include improvement of evaluation processes in the future?

It should be noted that Dressel suggests an additional step not included by Guba and Stufflebeam—an evaluation of the evaluation—asserting that evaluators must assume at least partial responsibility for unsuccessful evaluations. This point will be discussed further in the concluding section of this monograph.

The CSE Model

The final model that we will discuss is the decision-oriented model developed at UCLA's Center for the Study of Evaluation (CSE) and described by its former director Marvin Alkin. The foundation for the model is Alkin's (2) definition of evaluation:

> Evaluation is the process of ascertaining the decision areas of concern, selecting appropriate information, and collecting and analyzing information in order to report summary data useful to decision-makers in selecting among alternatives.

Because the definition, as well as the assumptions on which it is based, are closely tied to the decision-making process, evaluations are classified according to five decision categories and the kinds of information required for making the decisions. Alkin refers to these as evaluation need areas.

The first need area is called *systems assessment* and refers to evaluations that are necessary to provide information about the current status of the system. The difference between what is and what is desired represents a need and results in a statement of objectives written in terms of desired program outcomes. The second area, *program planning,* refers to information that will help the decision maker select a particular program that is likely to be effective in meeting the specified needs identified in the first stage. The function of the evaluator is to provide information concerning the potential effectiveness of different courses of action so that decision makers can choose the best from among the alternatives presented.

Once the program has been selected (or designed), an evaluation of *program implementation* provides information concerning the extent to which the program is being carried out in the way it was intended and information showing whether or not it is being provided to the group for which it was intended in the program plan. *Program improvement,* a fourth need area similar to formative evaluation, requires evaluative information concerning the manner in which the program is functioning—the attainment of *en route* objectives, the presence of unanticipated outcomes, and the relative success of the different parts of the program. Information collected in this stage should include data on the extent to which the program is achieving its intended objectives and information concerning the impact of the program on other processes and programs.

The fifth and final area of the CSE model is *program certification.* Similar in concept to summative evaluation, the evaluator's function is to provide information concerning the worth of the overall program, again in terms of both the extent to which the objectives have been attained and the program's impact on the outcomes of other programs. The information collected by the evaluator at this stage should enable the decision maker to make decisions regarding the future of the program. As in the CIPP model, the decision maker has four choices: to retain the program as is, modify it, disseminate it or terminate it.

Stages two through five are similar to the first four stages of the Discrepancy Model, and the first two and the fifth stages are similar to the CIPP model's context, input, and product evaluations. Process, as defined in the CIPP model, has been separated into program implementation and program improvement, and as far as Alkin is concerned, cost-benefit analysis, the fourth stage of the Discrepancy Model, is assumed to be part of every stage in his model.

The advantage of the CSE Model is that it is applicable to the evaluation of both discrete, definable instructional programs and broad-scale educational systems. In fact, Alkin argues that evaluations at the macro level of large educational systems require total examination beyond determining the extent to which program objectives have been achieved. For large-scale evaluations, the examination must include inputs, descriptions of alternative processes used within the system, descriptions of the input-output relationship and data on unanticipated outcomes or consequences in addition to data on the achievement of intended or desired objectives. Unfortunately, Alkin's advice has not often been heeded.

Some New Approaches

Although not exactly models in the strictest sense of the word, the Modus Operandi Method and the Adversary Approach to evaluation must be mentioned, even if briefly, since they will both no doubt receive greater attention in the near future.

The Modus Operandi (MO) Method is suggested by Scriven (74) as an alternative when experimental or quasi-experimental designs cannot be used. The theoretical base of the MO method, which derives from procedures employed by historians, detectives, anthropologists, and engineering "troubleshooters," is really quite simple. A program is investigated to see if it was the cause of a certain set of effects. As Scriven explains, "the MO of a particular cause is an associated configuration of events, processes, or properties, usually in time sequences, which can often be described as the characteristic causal chain (or certain distinctive features of this chain) connecting the cause with the effect."

Certain effects are assumed to be caused by one or more factors, which Scriven calls a "quasi-exhaustive causal list." The presence of each of these factors is checked, and if only one is present, the investigator checks for a "causal chain"—the configuration of characteristic events, processes, or properties that may connect the cause with the effect. If one causal chain is present, that chain (not the butler) is the cause. If more than one complete chain is present, the possible causes associated with it are considered co-causes.

Although Scriven suggests using the MO method in situations where classical design cannot be used, he also argues that even in experimental studies some attention should be given to the questions implicit in the MO approach: "What are the means whereby the putative cause is supposed to be bringing about the effect? What are the links in the causal chain between them? Can we look for these links or arrange that they will be easy to look for? Can we use their occurrence to distinguish between the alternative causal hypotheses? How?"

The MO method is still in a theoretical stage and has not been tested in actual evaluation practice. However, it offers evaluators a logical alternative to employ in appropriate situations, and in line with Scriven's other contributions, could ultimately prove useful.

The Adversary Approach offers less promise, at least according to some who have used it in practice—for example, Popham and Carlson (62). First suggested by Guba (35), the Adversary Model derives its origins from the legal model of advocate/adversary conflict, and confrontation and third-party resolution. Although there are several variations in the actual way it is applied to evaluation (and the reader is urged to consult the several descriptions of the approach),[5] Adversarial Evaluation basically involves two separate evaluation teams (or individuals)—one chosen to represent the program in question and gather evidence in its favor; the other to represent a competing program, or, in the absence of a competing program, to gather evidence and present a case *against* the program. The results of the two evaluations are presented either in written reports or in a traditional debate setting, with the decision makers rendering the final verdict.

In theory, the Adversary Model seems to be an ideal way in which to be assured of a truly objective evaluation, and its champions extoll this virtue. But, according to Popham and Carlson (62), the model has several serious defects: it is dependent upon the two competing evaluation teams having equal skills and on the commitment and fairness of the "judges;" there is no adversary court of appeals to which an improper ruling can be protested; it is expensive; and lastly, most educational decisions are not amenable to the binary choice of a winner/loser or go/no-go adversary contest. Educational decision makers need many more options concerning the future of a program than just those of maintenance or termination. The ultimate fate of the Adversary Model will have to await more reports of its use in actual evaluations. Perhaps when guidelines for its use are refined, some of the deficiencies encountered by Popham and Carlson will be remedied.

Citations

Countenance Model—Stake (83)
Differential Evaluation Model—Tripodi, Fellin, and Epstein (94)
Priority Decision Model—Boyle (9)
Trade-off and Comparative Cost Model—Glass (30)

---

[5] See Guttentag, M. (38); Kourilsky, M. (46); Levine, M. (47); Owens, T. (56); Wolf, R. L. (109); and Wolf, Potter and Baxter (110).

Systems Approach Model—Yost and Monnin (112)
Cost Utility Model (Costa, 1973)

Ontological Models—Peper (58)
Synergistic Models—Hunter and Schooley (43)
Ethnographic Models—Dobbert and Dobbert (20); Wilson *et al.* (108)

EVALUATION DESIGNS

The evaluation models described in the previous section represent the major paradigms of educational program evaluation; they have been used to guide many evaluations and they have influenced the thinking of many practicing evaluators. Models provide a broad base for designing evaluation activities by offering a framework and conceptualization that guides both the focus of the evaluator and the orientation of the evaluation. But models do not provide strategies for implementation. "Although models may help the evaluator isolate the types of decisions to be made, they do not provide procedural guidelines regarding *how* those decisions should be made." (60) Guidelines are provided by the design, which establishes the conditions and procedures for collecting the data required to answer the questions of concern. The design must be related to the type of program or service being evaluated; that is, the selection of a particular design is guided by the decisions that will have to be made as a consequence of the data. In turn, the adequacy of a particular design can be determined by the extent to which the results may be interpreted and the questions answered. In most cases, evaluation designs have been borrowed from research.

For example, Campbell and Stanley (12) distinguish between three types of research designs commonly used in evaluation—pre-experimental, experimental, and quasi-experimental—evaluating a number of specific designs in each category according to their ability to withstand threats to their validity. That is, the criterion differentiating the three groups of designs, as well as the quality of the designs within each group, is the extent to which the design protects against the effects of extraneous or nonprogram variables, thus legitimizing the results that are attributable to the program. More specifically, the criterion is the extent to which the design protects against eight threats to internal validity[6]—eight kinds of variables, extraneous to the program, that if not controlled, will affect the outcomes of the program and thus the accuracy of the interpretations that can be made of the data.

---

[6] Campbell and Stanley also describe threats to external validity that jeopardize the generalizability of the findings. Although some writers argue that generalizability is (or should be) an important consideration in program evaluation, most others feel as we do, that generalizability is not a major concern in most educational program evaluations. For a description of threats to external validity, the reader is referred to Campbell and Stanley (12).

The eight threats to internal validity are as follows:

History: Outside events, such as changes in factors like the job market, the economy, or television programming, can affect the subjects of a program and thus the program results. Outside events are likely to occur when the program being evaluated extends over a long period of time.

Maturation: Processes within respondents, such as fatigue or growth, produce change as a function of the passage of time. Natural growth alone may sometimes be responsible for changes that are observed in a program evaluation. Weiss (99) describes the problems confronted in evaluations of delinquency prevention programs that do not have control groups. Because young males generally become less likely to commit crimes and more likely to hold jobs around the age of 17 or 18, when such results appear in program evaluations, they cannot be attributed to the prevention program unless a control group has been incorporated in the design.

Testing: The effect of a test on the scores of a second test, as in the pretest-posttest design, prevents a true determination of the program results.

Instrumentation: Changes in the instruments themselves, in calibration or difficulty level, or changes in the observers or scorers used affect the accuracy of interpretations.

Selection: Biases resulting from the differential recruitment of the experimental and control groups affect the accuracy of interpretations.

Statistical Regression: Non-program effects can appear during statistical manipulations. When groups are selected for a study on the basis of extremely high (or, more often, low) scores, their scores on subsequent tests will tend to regress statistically—that is, move back toward the mean of the group. The regression is an artifact of the statistics and not an effect of the program.

Selection-Maturation Interaction: Selection biases result in differential rates of maturation or changes as a function of time.

True experimental designs protect against all of these possible threats to internal validity; quasi-experimental designs generally protect against most of them. Quasi-experimental designs require the same rigor, but they are more practical than the true experimental model in many real-world situations. Pre-experimental designs totally lack control and, according to Campbell and Stanley are "of almost no scientific value." Examples of pre-experimental designs are: 1) the one-group, pretest-posttest design in which a single group is pretested, exposed to a program, and then posttested; depending upon the length of time between the pretest and posttest, the design is open to the threats of history or maturation; 2) the static-group comparison, in which a group that has received a program or service is compared with a group that has not—a comparison that is suspect since the original equivalence of the two groups is unknown; and 3) the one-shot case study in which a single group is studied once. More will be said about the limitations of case studies in a subsequent section of this chapter.

Quasi-experimental Designs

Because of the difficulty of conducting true experiments in the real world of education, quasi-experimental designs have become more widely used in both research and evaluation projects in recent years, particularly as these designs gained respect under Campbell and Stanley's sponsorship. The designs described on the following pages are the more widely known of the quasi-experimental group, and each claims certain special features that makes it appropriate in different types of evaluation settings. For a more exhaustive list and description of designs, the reader is referred to Campbell and Stanley (12).

The Nonequivalent Control Group Design: Probably the most commonly used design (and also the least satisfactory) is the nonequivalent control group design, in which control and experimental groups are formed without benefit of random assignment. A comparison group of available individuals or intact groups whose characteristics are similar to the experimental group are used as controls. Pretest and posttest measures are taken for both groups and the results are compared. Although obviously not as rigorous a design as a true experiment, in which comparison groups are based on random assignment, the main issue in the nonequivalent control group design is one of selection—identifying the variables that were used to place the participants in each group. The objective, of course, is to make the two groups as similar as possible. The more similar the control group is to the experimental group, the more reliable the interpretations that can be made of the data. Popham (60) and Weiss (99) both provide suggestions for increasing the similarity of the two groups. Popham suggests reviewing the scores of the pretest for both groups and selectively eliminating the "discordant learners" from the posttest analysis. Weiss proposes using "unawares" (people who did not hear of the program but might have joined if they had) and "geographical ineligibles" (people with characteristics similar to the experimental group who live in locations that have no similar program).

The Time Series Design: The time series design involves studying the behavior of an individual or a group over time. Although the statistical procedures for analyzing the data are sometimes complex, the time series design has many advantages to offer. A series of measurements are taken of the participants before, during, and after the onset of a program, with the before measures establishing a baseline performance level against which to measure changes. The measures are examined to determine an "effect pattern" or trend to show the impact of the program over time.

The multiple time series design provides more rigor by adding an additional group and examining the series of measurements for both groups. If the program evaluated has been effective, the effect pattern for the two groups should be markedly different. A major advantage of the time series design is that it is a fairly powerful design, providing excellent information on the effects of a program even when a comparison or control group cannot be used. Time series designs are particularly well suited for longitudinal evaluations and social action evaluations where the program cannot be withheld from appropriate participants.

Experimental Designs

Although some writers acknowledge the difficulty of applying controlled experiments to the problems of education, and more than a few add the caveat of "where conditions allow," experimental design is to many educators the cornerstone of evaluation—the ideal methodology for educational program evaluation.[7] Campbell and Stanley (12) state unequivocally that they are

> . . .committed to the experiment: as the only means for settling disputes regarding educational practices, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties.

Classic experimental design incorporates two important techniques that together rule out the possibility that something other than the program caused the observed results, and thus, they confirm the legitimacy of the interpretations made from the data. These techniques are the use of control or comparison groups and randomization. Quite simply, this means that samples of the target population are randomly selected and assigned to either the experimental group receiving the treatment (program) or the control group, which receives a different treatment or no treatment. Members of the two groups are posttested after the program has been completed, the differences are compared, and the experimental program is pronounced a success if the experimental group has more of whatever the criterion variable is than the members of the control group. That the experimental group had fewer cavities after using Crest should by now be a familiar slogan.

---

[7] See Aronson and Sherwood (4); Campbell (10); Evans (26); Glennan (32); Houston (42); Popham (60); Porter (63); Rossi (72); Scriven (74, 76, 77); Stanley (86, 87); Welch and Walberg (102); Wholey *et al.* (107); and Weiss (98, 99). Evans (26) makes a compelling argument in favor of small-scale controlled experiments to test the relative effectiveness of alternative program techniques as a precursor to the introduction of massive national programs.

The essential feature in experimental designs is randomization, which increases the probability that subjects who form the control group are basically equivalent to those in the experimental group. In the Crest experiment, this meant that the people who formed the control group and used Brand X were, as far as the experiment was concerned, no different from the people in the experimental group using Crest—at least not until they completed the program. Controlled experiments reduce the possibility that something other than the program caused the results. Suppose, for example, that subjects were not randomly assigned to the Crest and Brand X groups and it turned out that the subjects in the Crest group lived in a community that introduced fluoridation into the water soon after the study had begun. Suppose, at the same time, that the majority of Brand X subjects lived in a community that did not have fluoridated water. Quite obviously, the inference that the continued use of Crest results in fewer cavities would have been suspect, and Arthur O'Connell would have been out of a job.

Without question, experimental design can be a powerful tool. If people can be randomly assigned and if there are enough of them available to form an experimental and a control group; if the control group will not be harmed or deprived psychologically, socially, or financially by not receiving the program or by receiving a placebo program; if the program is a specific, definable entity; and if the objectives are explicit, then an experimental design is probably the best choice. If the evaluation proceeds smoothly and if the instruments and measures are valid and reliable and appropriate to the objectives, then, if the experimental group shows greater positive change than the controls, we can be fairly certain that the change is due to the effect of the program.

But programs do not exist in apolitical or ideal contexts and compromises in design are inevitable. There are innumerable occasions when forming control groups and randomization are difficult; there are many situations in which it is impossible. Sometimes programs have to be offered to intact groups, such as classrooms already formed according to school schedules. Sometimes groups available for comparison are too dissimilar. Greenbert (33) and Weiss (98, 99) both comment on the problems associated with finding truly equivalent groups or communities where randomization has been possible and note that the alternative usually used, that of matching, is not a satisfactory solution. For every factor on which groups are matched, there are other equally, if not more important, variables on which they are unmatched. It is these variables that may in fact exert more influence on the outcomes than the variables on which the groups are supposedly matched.

In other situations, programs must be provided on a voluntary basis and made available to all who apply. This is particularly true in the case of social action programs whose primary purpose is to shift the position of a specified target group relative to the rest of society. Few administrators, or program evaluators for that matter, would be willing to deprive people of programs that would be of benefit to them. As Suchman (92) comments, it is difficult both to refuse service to those who seek it and to force it upon those who don't want it.

But, even when control groups are feasible, there are a number of problems that interfere with the operations of an experimental design. First, experimental designs are particularly vulnerable to "Hawthorne effects."* Regardless of randomization, the results of a program can become contaminated if either the experimental or the control group find out that they are participating in a "study" and become aware of their special status. Experimental participants may try harder while their control group counterparts may become annoyed or angry at being rejected by the program. The change in their actions or attitudes will affect the outcomes of the program. In addition, it is difficult to maintain contact with controls who are not receiving an alternative or placebo program.

An added problem concerns contamination of the control group. Mann (51) observed that in an organizational setting, innovations sometime "spread like a disease" to control groups. Rossi (72) notes in addition that a changing economic or political climate can make available to the controls programs or services that are essentially equivalent in many respects to the program or services being evaluated. It is far easier to implement a rigid evaluation in programs operating in highly centralized organizations such as prisons, hospitals, or boarding schools in which the organization maintains strict control over its members and the evaluator can thus maintain strict control over the design.

Still, as Weiss (99) clearly points out, ingenious adaptations can be made to alleviate, and in many cases eliminate, most of the problems that beset experimental design. Scriven (76) suggests the use of multiple experimental groups to separate Hawthorne effects from those of the programs. Weiss and others (12, 24, 44, 60) suggest the time series design in which the treatment group becomes its own control through repeated measures of outcome variables or in which two different programs are compared and the treatment group of one program serves as the control group for the other and vice versa. Rossi (72) proposes a two-stage evaluation consisting of a reconnaisance phase in which non-experimental designs are used to screen out programs that should (and can) by investigated further and an experimental phase in which powerful controlled experiments are used to evaluate the differential effectiveness of a variety of programs that demonstrated sizable effects in the first phase.

---

*The term refers to a series of studies made at the Hawthorne Works of the Western Electric Company between 1927 and 1932. Researchers found that workers increased production whenever they became the subject of attention in a study. The "Hawthorne effect" has subsequently been found in many research and evaluation situations where experimental designs have been used.

The experimental model has been challenged not only because of the inherent difficulties in using such designs but also because in many instances experimental designs are counter-productive to the needs and goals of the evaluation. As we pointed out earlier, the design must be suited to the purposes of the evaluation. If the purpose of an evaluation is to find out how well a particular program achieved its goals, an experimental design is ideal. If decision makers are concerned with program implementation, participant satisfaction, or information for program improvement, other designs are far more appropriate. In these examples, experimental design would be inadequate for the task.

The many limitations of experimental design, particularly those which focus on the extent to which a program has achieved its objectives, are well documented and will not be reiterated here. For more detailed discussions, the reader is referred to Borich and Drezek (8); Guba (34); Riecken (66); Rose and Nyre (71); Stake (84); and Wergin (103).

Most studies carried out under experimental conditions fail to assess the impact of the program operating within functioning institutional or organizational systems. The focus on objectives limits the evaluator's understanding of the program and, despite Scriven's exhortations, attention is seldom paid to the merit of the goals established for the program or to unanticipated outcomes that may have far more important consequences than the goals originally intended. An obvious example is a math program that significantly improves children's understanding of mathematics but results also in their hating math! Experimental designs do not take into account changes in goals (or procedures) that frequently take place once a program is underway, and they cannot provide the immediate formative feedback that programs often need in order to identify and correct snags in their early stages of implementation.

House (41) offers an interesting analysis of the problem, arguing that the classical approach to program evaluation, in which learner performance is measured on standardized tests of achievement (which implies that the larger the gain, the better the program), is based on utilitarian ethics. Utilitarian ethics stipulate that a society is just when its institutions are arranged so as to achieve the greatest net balance of satisfaction as summed over all individuals. The principle of utility is to maximize the *net* balance of satisfaction. Thus, a common measure or index of the criterion is required so that quantitative calculations can be made. In education, that measure is the standardized test, and in the classic evaluation approach, the best educational programs are those which produce the greatest gains in test scores regardless of the distribution of those scores. Only the final, net score counts, and, since it is averaged across all individuals, one person's loss is balanced by another person's gain. The real effect of the program on different subsets of individuals is masked.

Most experimental designs that have been used in educational evaluation fail to consider the manner in which the program was implemented or the configuration of people, events, processes and practices, values and attitudes that surround the program, affecting the environment in which it operates and thus, at least presumably, its outcomes. It is not enough to document that a program failed to work. It is essential to identify the processes and other variables that combined to defeat it. Particularly in the case of large social action programs, but even with small-scale educational programs, the investigation of negative effects is an important issue. The capacity of communities, organizations, institutions (and people!) to resist change must be investigated and the factors that defeated a program identified so that they can be used as a base for the design of a program that is more likely to be effective.

Conversely, it is not enough to document that a program achieved its goals and the extent that it did so. Equally important as the attainment of goals is the concern with *why* the results occurred, what processes intervened between input and outcome, how the program actually operated, what non-program events may have affected participation, and what implications and guidelines can be derived from the evaluation for program improvement and replication. Experimental design alone cannot provide this essential information.

Weiss and Rein (101) point out that in broad-aim programs, different approaches are often used at the local level so that the programs in effect differ from community to community. A description of the different forms and approaches as well as the forces that shaped each would be important information that cannot be obtained through traditional experimental evaluation.

Stufflebeam (89) contends that experimental designs are only appropriate in product evaluations and, thus, are of minor relevance to educational evaluation. Guba (34) goes further, stating that experimental design actually "*prevents* rather than promotes changes" because the programs cannot be altered if the data and interpretations about the differences between them are to be unequivocal.

The same criticisms and shortcomings can be leveled against quasi-experimental designs in which the usual thrust of the study is also the degree to which desired goals have been attained. No matter how effective and useful they are in some situations, again, little attention is paid to how the program developed, what unanticipated consequences occurred, what variations exist among the program's component parts or units, what outside events affected either programming or participants, or to the adequacy of the program operation and the capability of the staff. As Stake (83) suggests, most classical designs were developed as a means of examining "minute details"; they were not developed for portraying the "whole cloth of the program". The point is, evaluation designs must accommodate the characteristics and informational needs of the program, not the other way around.

Process Evaluation—the Other Extreme

Unfortunately, the very real problems with experimental designs and the deficiencies of quantitatively oriented evaluations that reached their height in the era of accountability precipitated a reactionary movement to the other extreme—an equally deficient process-oriented approach, alternately referred to in the literature as transaction-observation, process-oriented, qualitative, or illuminative evaluation.[8] These approaches, which derive primarily from Stake's countenance model and his later "responsive" evaluation, focus almost exclusively on the environment or "milieu," eschewing quantitative output measures, and are preoccupied with program process.[9] Nonexperimental designs (pre-experimental in Campbell and Stanley's terms), which were previously considered to be of little or no value to educational evaluation—at most, a last resort—have suddenly come to be the method of choice (49, 57, 79, 82). Most popular is the case study, in which the evaluators "observe, inquire further and then seek to explain" (57). The data base relies heavily on interviews and observations, often informal. The evaluator documents and describes what it is like to participate in the program, how participants feel about the program and the staff, how the staff feels about the program and the participants, and what both parties believe to be the significant features of the program. Surrounding elements of the organization and environment are investigated and their relationship to the program is explored. Anecdotes are collected and program documents are reviewed. But the whole issue of program outcomes—the consequences of a program—is totally ignored.

A goal-attainment model that excludes process data can only address the issue of what has happened. It cannot respond to the broader quesiton of what was responsible for which outcome. Even more important, it cannot provide information for program improvement and development. The process-focused approach, which excludes outcome data, cannot deal with either question.

In their extensive critical review of federally-sponsored evaluations, Bernstein and Freeman (6) comment on a study whose data analysis techniques included reviews of narrative descriptive reports and impressionistic summaries obtained by means of the case-study approach as follows:

---

[8] Although Parlett and Hamilton have popularized the term illuminative evaluation, credit for coining the phrase and suggesting the methodology and issues of concern belongs to Martin Trow, who spoke of the need for illuminative evaluation in 1970.

[9] Process as used here is a broader concept than the traditional one where process evaluation means to determine whether or not a particular program was implemented according to its plan and directed at the appropriate specified target population. As used here, program process refers to the resources and forces external to the program that may affect its operations and includes, in addition to the above, an investigation of other programs and components within the institution and the needs, resources, and attitudes of the larger community.

> We cannot avoid noting that this study indicated . . . that no measures of outcome were taken at all. Barring some very unusual circumstances, we would conclude that this study is illustrative of an evaluation which did not meet the basic requirements necessary to be classified as competent evaluation.

Sadly, this approach is particularly appealing to the fainthearted. Because it typically eschews making judgments about the worth of a program (probably a wise decision in view of its lack of rigor), this approach is obviously tempting for those who wish to avoid the risk of finding their programs impotent. All they have to do is ask participants how they felt about a program, chronicle how the administrators and the staff felt, describe the institution and the program, write up an interesting narrative report, and ignore the fact that no matter how richly evocative or interesting the report, the findings may well be distorted and untrustworthy.

Using somewhat different terms, Scriven (76) differentiates between approaches to educational evaluation in which the emphasis is on intrinsic criteria and approaches in which the chief attention is given to extrinsic criteria. Intrinsic criteria refer to the constitution, nature, or essence—the qualities inherent in the subject of evaluation—and are associated with its process. Extrinsic criteria are concerned with the effects of the program. Both Scriven and Popham (60) argue that the emphasis on intrinsic criteria is all too common in educational evaluation, and that most such studies are too haphazard to be properly considered systematic evaluations.

Case study evaluations are seriously defective in a number of ways. At best, they are vulnerable to the threats of history, maturation, selection, and mortality. Because there is no design directing the data collection or guidelines that establish parameters, case studies accumulate a huge bulk of data, much of which is irrelevant and all of which is difficult to organize (101, 103). And, of course, there are no baseline measurements with which to determine change or growth. But far more serious are the problems of bias and subjectivity that are endemic to the case-study approach.

Case studies operate within relatively small units of analysis, and assessing a program by judging only a few units exposes the study immediately to sampling bias. There is great variation in the reports provided by interviewees because of their biases, and this phenomenon is not eliminated by "triangulation." A key concept in many case study methodologies, triangulation is a term borrowed from Webb *et al.* (97) that refers to viewing the problem from a number of angles and representing the perceptions of the program by its different publics to ensure a fair evaluation. Areas of agreement and conflict are identified and defined as the evaluator attempts to find convergence of findings from a number of different sources (57). The problem is that the perspective of a given "public" depends entirely upon which members of that public are interviewed and what they are willing to tell. There may well be, in fact, several different perspectives within a particular public, and unless the interviewees are selected randomly, the story they tell may well represent the biased view of only a few members or at most one faction of the group. Riecken (66) attacks not only the sampling bias in case studies, but also the lack of comparability between subjects' reports, which severely limits statements of the extent to which particular effects were produced.

Case study process evaluations will always be vulnerable to charges of bias either on the part of the participants or the evaluator. Lucco (48) goes so far as to question the "political underpinnings" of evaluations in which emphasis is placed on program operations and process. In order to bring a semblance of quality control to a case study, the investigator must be conscientious, skilled, insightful, and objective. But, even where evaluators are paragons of brilliance, objectivity, and virtue, their observations are still made from their personal frame of reference, and subjective bias is impossible to avoid.

House (41) and Stake (82) both attempt to justify the subjective nature of case study evaluations by comparing the procedures to those of an anthropologist or historian. An anthropologist observes a tribe or village in order to describe its culture, the roles and relationships of the members, and the way in which it functions. Historians describe events in order to identify patterns and causal relationships between events. But anthropologists and historians are interested in describing and interpreting only; they do not make judgments nor do they need to make decisions. In education, we need to make decisions, and evaluation increases the rationality of these decisions. Evaluation always has a heavily subjective component because it deals with values; but that does not in itself excuse slovenly design or statistical analysis (21). The intent should always be to move as far as possible toward objectivity and clarity. Illumination is not evaluation.

Still, there are some situations in which the evaluator simply must use limited methodological tools. Certainly it is better to know how faculty and students, or any subjects of a program for that matter, behave while they are under observation than to know nothing at all about how they behave. Weiss and Rein (101) suggest that informal approaches usually associated with exploratory research, such as the case study, may be appropriate where the relative contributions of various components of a large-scale program are difficult to determine because of the participants' uncontrolled exposure to the program or where it is difficult to select and operationalize evaluative criteria that are sufficiently broad in scope to reflect the program's full range of consequences. They also suggest that qualitative appraisals by means of case study can be used to describe the variations in social action programs from community to community in combination with an assessment of overall program outcomes through experimental design.

In these situations, observational techniques and interviewing can provide useful (and rapid) additional feedback. And, as an exploratory analysis, case-study data may provide the evaluator with suggestive leads concerning significant variables that can subsequently be studied more rigorously with an experimental design (103). Mann (52) is less sanguine, however, suggesting that these leads may be suspect in light of the tremendous bias implicit in the case-study approach.

As with experimental design, of course, the process-oriented case-study approach has its own band of loyal followers for whom case study is *the* method, enabling the evaluator to understand the whole of a program through direct and vicarious experience. And without question, it is important to understand the "whole" of the program—including the differences in perspectives between program planners and program operators, differences in values and perspectives of different audiences, ways in which the program operates, and other programs, people, events, or combinations thereof that may influence the program under analysis. Understanding what happens with respect to the political and social forces involved is essential if a program is to address the issues or problems effectively. Few professionals would deny that an understanding of process is important. One has only to look at the legal profession, where the integrity of the process by which one is brought to trial dictates the outcome. But, as Weiss (99) argues, critical as it is to learn more about the process and dynamics of a program, it is nevertheless equally critical to determine its outcomes.

Identifying the outcomes of a program is only part of an evaluator's task. Unless an evaluation describes the actual program and the procedures and processes that brought about the outcomes, it is presenting a half-told story.

But, understanding the process without defining the outcomes is also an unfinished story. A recent "human" interest story reported in the *Los Angeles Times* (Monday, July 4, 1977) provides an amusing illustration of a process-only orientation. The story was about an operation to reset an elephant's broken leg in New Delhi. "The operation was successful; the operated limb was corrected," claimed the team of veterinarians. And they went on to describe how an army tank crane, 12-inch steel pins, welding equipment, yards of plaster of paris and gallons of antibiotics were used in the surgery. The fact that the elephant died of heart failure caused by nervousness and excitement during attempts to get her on her feet during postoperative procedures did not stop the veterinarians from stressing the success of the surgery and was only peripherally noted. The story headline read "Operation Success but Elephant Dies"!

Clearly, both the process-only and the outcome-only approaches are inadequate for the evaluation of educational and social programs. What is needed is a methodology that combines rigorous experimental data with "a natural history account of events and actors before, during and after program implementation" (5). Integrated evaluation approaches such as the ones described in the next chapter may well provide the answer.

INTEGRATED APPROACHES TO PROGRAM EVALUATION

The concept of integrated evaluation is not new. As far back as 1963, Cronbach stressed the need for evaluating the interactive events or "process" of the classroom in addition to the learning outcomes. Scriven (76), too, offered what he called mediated evaluations, which combined attention to both intrinsic and extrinsic criteria. Suchman (92) proposed four different kinds of evaluation: evaluation of effort, or the amount of action involved in establishing a program; evaluation of effects, or the results of the action; evaluation of process, the way in which the effects were achieved; and evaluation of efficiency, the ratio of costs to effects. And, in Stake's Countenance Model described earlier, transactions are equivalent to process.

An integrated evaluation approach is a hybrid of the two polar positions described in the last chapter, one that combines the study of program process with the study of outcomes. In this section, we will focus on two examples of integrated approaches—holistic evaluation and transactional evaluation. Brief descriptions of these programs will be followed by examples of actual evaluations in which these approaches were used.

Holistic Evaluation

Holistic evaluation is an integrated, multidisciplinary approach to program evaluation that investigates *both* process and product (45, 71). By broadening the paradigm, holistic evaluation enlarges the scope of questions that can be asked and the body of data that can be collected. It includes descriptions *and* quantification, objective data *and* perceptual reports, and it can accommodate experimental designs as well as case studies. Named to convey its sense of comprehensiveness—not its "holiness"—holistic evaluation rests on six basic assumptions:

1. Programs do not exist in isolation. Educational and social programs are but one component within a broad system or organization in which program activities are carried out.
2. As such, programs receive influences from various people and groups with differing needs, interests, and points of view.
3. Educational and social programs have different meanings and different implications for these different groups.
4. The evaluation of these programs involves gathering information useful to the disparate groups of decision makers with direct input into the program as well as groups which may not be directly involved in the

program but whose decisions may nevertheless affect it.

5. Procedures for carrying out the evaluation must be appropriate to the program and selected to provide the kinds of information that are required by the different groups of decision makers. (In other words, the decision-needs dictate the methodology of the evaluation.)

6. Most decisions, by their nature, require information about both program process and program outcomes.

Holistic evaluations are thus concerned with four major areas: 1) the social-psychological environment in which the program operates; 2) attitudes, values, interests, and perceptions of participants and surrounding groups; 3) program and participant outcomes; and 4) the interaction of the various elements comprising the system that may affect the operation of the program, and thus its outcomes.

Holistic evaluation is not a model in the strict sense of the word. It is a conceptual framework with certain defined strategies from which program-specific (and site-specific) procedures can be derived for either formative or summative evaluations. Holistic evaluation has been used to evaluate four federally funded programs in vocational education (45); a multi-campus instructional development program for faculty (70); a statewide program operating in three public segments of postsecondary education (69); a curricular program at a professional school (55); and a statewide program for disadvantaged students (27). The last two evaluations will be described in depth in the section of case studies to illustrate the holistic approach.

Transactional Evaluation

Transactional evaluation is a term usually credited to Robert M. Rippey. According to Rippey (67), the actual meaning of the term is still emerging; it is not yet fully developed. A synthesis of the writings of several transactional evaluators, however, shows that transactional evaluation has certain attributes that distinguish it from so-called traditional approaches to evaluation.[10]

To begin with, transactional evaluation emphasizes a broad base of participation. It involves not only the designers and supporters of a program, but also a representative sample of antagonists—persons who are likely to be affected adversely by the program or disturbed by the consequences of change. Secondly, transactional evaluation stresses the value of conflict and uses it as a basis for examining differences in perception among the various groups. In transactional evaluation, the key is not consensus, but an exploration of the divergent views which result from different perceptions and an examination of their implications for decision making. All new programs create some dysfunction in existing school/community relationships. In transactional evaluation, changes resulting from the creation or addition of a program are continuously observed and resulting conflicts are brought to the surface.

---

[10] The reader is referred to Rhine's (65) case study of the longitudinal evaluation of Follow Through, which provides a good example of the distinguishing characteristics of transactional evaluation.

A third part of transactional evaluation is the Transactional Evaluation Instrument—both a product and a process that permits protagonists and antagonists to clarify their perceptions and uncover sources of conflict or perceptions of conflicts that were submerged.

Finally, transactional evaluation differs from traditional approaches in the emphasis it places on diagnosis and improvement rather than on establishing the superiority of one program or method over another. Although, again, there is some disagreement among writers, Scriven insists upon the importance of designing evaluations as comparative experiments on the ground that judgments of worth *are* comparative (31, 76). Transactional evaluation is not concerned with comparative worth; it is concerned with social and organizational relationships. According to Rippey (67), the key to the transactional model's effectiveness "is the continuous evaluation by both protagonists and antagonists, of both the expected and unexpected consequences of change" in order to modify and improve the program.

Grounded in organizational theory, the function most suitable to transactional evaluation seems to be the evaluation of institutional change projects. As Rippey acknowledges, transactional evaluation is based on "a study of internal conflict concomitant to change." Rippey includes transactional evaluation as an essential step in a change strategy which proceeds first to establish disequilibrium; increase differentiation; begin change on a small scale under the best possible conditions (which Rippey later explains as first working only with those who support the change); improve the climate and organizational mechanism for change; and lastly, implement all new programs as temporary, small scale, pilot experiments so that the effects can be studied without undo disruption to the entire social organization. Transactional evaluation requires that protagonists and antagonists jointly establish the criteria for assessing and measuring both the planned and unplanned outcomes.

Transactional evaluation consists of two main stages. In the first stage, the transactional evaluator aims to uncover the sources of conflict; in the second stage, the evaluator uses both proponents and opponents to develop the evaluation plan. In order of sequence, transactional evaluation proceeds as follows.

First, all of the groups involved in or likely to be affected (directly or indirectly) by the change (program) come together for a series of meetings. Three conditions must be met during this first stage: 1) all groups affected directly or indirectly should be represented; 2) a neutral party should conduct the meetings; and 3) sessions should be conducted in a nonjudgmental manner. Although feelings of suspicion and distrust are prevalent, the issues and sources of unrest may not be clearly defined and the problems may not necessarily be those that are articulated. But the climate thus created is a necessary condition for the subsequent development of the evaluation plan.

The second stage involves construction of the transactional evaluation instrument, the key to unlocking conflicts and controversies. Again, everyone is involved in the process. The evaluator first formulates a general statement of the issue in the form of a question based on the feelings expressed in the initial meetings. Each participant in the group is then asked to respond to the question with a series of statements. These responses are collected, tabulated, and categorized with the original wording retained wherever possible. These responses, in effect, become the items for the instrument.

The transactional evaluation instrument is administered, and participants respond to each of the items appropriate to their role group (for example, teachers, administrators, students, parents, and so forth). Responses are tabulated, a master copy is prepared, and copies are distributed to participants. Finally, the last and most important step is the examination of responses, which reveals the areas of shared values and goals and the areas of open conflict.

In the second phase of transactional evaluation, the proponents and opponents of the program (or a particular aspect of the program) develop and implement an evaluation plan with technical assistance provided by the professional evaluator, who, according to several transactional writers, should be a fully participating member of the program staff. The presence of both those who are for and those who are against the program insures that program monitoring includes not only the outcomes intended by the proponents but unexpected negative outcomes suggested by the opponents. Nonbelievers who are apprehensive about their roles once the new program is implemented can often be reassured by direct action of the project, inservice training where necessary, or clarification of policy. But even more important, initial opponents can be given a legitimate role in the program, one that often leads to their conversion and ultimate support, or, at the least, their understanding and tacit agreement. Resistance may be identified and dealt with at each stage of the process of change—when the innovation is initiated, when it is being evaluated, when the findings of the evaluation are accepted, and when further changes in the program are recommended.

The insistence upon the involvement of both factions rests upon George Simmel's (78) theories of working relations. Simmel argues that the basis for a positive working relationship is an interaction in which both parties have parity in the exchange; where the relationship is not reciprocal, one party is diminished and becomes dissatisfied in the relationship.

Transactional evaluation is not as suitable for large-scale, summative evaluations, although Cicirelli (15) suggests that even a large-scale summative evaluation, such as that of the Head Start Program for disadvantaged children, may be made more effective if the two major principles of transactional evaluation are incorporated into the evaluation; that is, the groups that might feel threatened or adversely affected by the program (or the evaluation) and thus resist it are identified, and representative samples of these groups are involved in the evaluation from the planning stages through the implementation stage and during consideration of findings and implications.

Transactional evaluation is similar in many respects to formative evaluation, particularly in its concern for continuous diagnosis and program improvement. But transactional evaluation broadens the scope of formative evaluation by involving a larger group of individuals, eliciting a wider range of opinions and values, and giving more continuous attention to information concerning the institutional role. When a program of change looks beyond the immediate outcomes of its intended goals, examines the roles and apprehensions of all parties to the system, and attempts to continuously monitor its total effects, that program is participating in holistic or transactional evaluation. Case studies of transactional evaluation are presented in the next section.

CASE STUDIES

The Evaluation of Social-Action Programs

The first two case studies described in this section concern evaluations of social-action programs—programs designed specifically to improve the life conditions of a particular group of people. These programs vary in scope—some cover the nation; some, a state or a city; and some are confined to a single site. Social-action programs also vary in size. Some serve thousands, others, hundreds, and still others serve a relatively small number of people. Some social-action programs are aimed at a clear-cut, single purpose, such as improving children's ability to read. Others are more complex and are aimed at alleviating a broad-based, pervasive social problem, such as programs designed to improve mental health or provide equal educational opportunity. These programs have at their base long-range goals which may or may not be attainable within the lifetime of any one evaluator. (Contrary to some thinking, evaluators are mortal and they have only one life in which to evaluate.)

Social action programs are rarely confined to a single focus. More often, program areas are legion, ranging from educational and social welfare to medical and legal services. The common thread that runs through these programs regardless of emphasis, however, is their goal of improving the life condition of the people they are intended to serve. Because of the magnitude of this goal, the vast sums of money that have been allocated in order to attain it, and the variety of services and programs offered, evaluation can play an important role in assuring that the programs serve the targeted population in the most effective way.

Understandably, decision makers, particularly legislators and governmental agencies charged with funding these programs, want to know if the expenditure is justified. Is the program meeting the goals for which it was established? Can it do so for less money? Should the program be expanded, reduced, eliminated? Can the program be more effective if it is revised? Regardless of the legitimacy of these questions or the sincerity of the questioner, however, these questions can be political and thus their answers politically loaded. In Cohen's (16) words, "Evaluating social action programs is only secondarily a scientific enterprise. First and foremost it is an effort to gain politically significant information about the consequences of political acts."

A most important issue for social action programs, in addition to overall worth, is program improvement. It is quite unlikely (probably more so for political reasons than for humane concerns) that any large-scale, broad-aim federal or state social program will be eliminated or even seriously limited as a result of any evaluation, no matter how inconsequential an effect the program appears to be having. What is more likely is that the results will be used to make the programs more effective and more responsive to the needs of those the program is serving. Why is the program not more effective? How can the services be improved? What other services should be added? These questions are far more important for the ultimate solution of the problems these programs were designed to address. Still, attention must be paid to both sets of questions within the context of each program site, taking into consideration local program variations. The problems in methodology as well as the constraints arising out of political and emotional factors are discussed in the following two case studies. The first case study illustrates the holistic evaluation approach; the second case study provides an example of transactional evaluation.

EOPS: A Case Study of Holistic Evaluation

Extended Opportunity Programs and Services (EOPS) is a special program established in the California Community Colleges for the purpose of providing equal educational opportunity to racial and ethnic groups and the minority of whites who had formerly been denied access to college because of deficient academic backgrounds and/or a history of poverty. In order to help these people gain access to college and meet the demands of academic life, the program provides financial aid and supportive services in the form of tutoring and counseling.

EOPS was conceived in response to the civil rights movement of the 1960s and the consequent political pressures to remedy the neglect of large groups of people by our major social institutions. As with the rest of the country, California's higher education system served primarily white, middle-class, economically advantaged students. Prior to the establishment of EOPS, a white middle-class student was twice as likely to enroll in college as was a member of a racial or ethnic minority. The wave of social consciousness that gave rise to the massive federal programs such as Project Head Start, Follow Through, and Title I of the ESEA also stimulated the thinking of California's leadership, and in 1968, Senate Bill 164 established the Extended Opportunity Programs and Services in the California Community Colleges.

Like so many other social action programs established at that time, EOPS expanded rapidly, growing from a $3 million program in 46 community colleges to a $7.6 million program in 94 community colleges in less than ten years. The speed with which EOPS escalated compounded many of its early deficiencies and added to the difficulty of its later evaluation. Staff were hastily selected without full attention to their qualifications as administrators. Programs were often instituted without adequate consideration of the goals and needs of individual colleges or the values and attitudes of the college and community membership. Many campus programs lacked careful planning. Participant data was seldom recorded, and few campuses documented the process of implementation. Other than the head-counting reports submitted annually to the Board of Governors, the policy-making body for the California Community Colleges, no systematic evaluation of the program was ever undertaken. However, the economic recession of the 1970s, coupled with the growing suspicion that massive social-action programs had not significantly alleviated the country's major social problems, finally led to a concern for evaluation, and in 1975 a "formative" evaluation of this multicampus program was conducted seven years after it began.

The Situation: The major purpose of the evaluation, as stipulated in the evaluation contract, was to determine the extent to which the community colleges had met the objectives of the legislation, those of the Board of Governors, and those of the individual colleges.

All told, there were 31 major objectives, ranging from those aimed purely at implementation (for example, "the community colleges shall establish . . .") to those aimed at various student outcomes. The objectives, like the program, were also seven years old, but it was clear that at least a part of the design would be simple. A straight accountability approach could be used to determine if the colleges did in fact do what they were supposed to do—establish financial aid and supportive services for persons of minority and/or disadvantaged background. But an additional charge of the contract was that specific recommendations be made regarding program improvement at both the state and local levels, and fulfilling this requirement was hardly a simple matter.[11]

Broad-aim social action programs are not one dimensional; rather, they are composed of a vast array of complex, interactive elements loosely called a program. If the purpose of an evaluation is, at least in part, to provide information for making improvements in a program or particular parts of a program, then it becomes necessary to distinguish the differential impacts of these parts and the processes that contributed to them. In the case of the EOPS evaluation, this task posed several methodological problems. For example, two of the EOPS objectives concerned improving minority students' self-concept and instilling in them pride in their cultural distinctiveness. Both of these objectives are noble, and they are plausible within the parameters of a social program designed to equalize educational opportunity. But to empirically distinguish the elements or parts of a program directly aimed at improving self-concept or instilling cultural pride necessitates not only a specification of criteria concerning what constitutes positive self-concept and cultural pride, but, even more, a knowledge of what in fact influences the development of these qualities. No one yet knows what educational or social practices or policies contribute to self-concept and cultural pride. We can speculate that a "supportive" environment (and this, too, needs clarification) may contribute to a feeling of acceptance, which, in turn, might enhance self-concept. But in the absence of a specific program designed especially for improving self-concept which can be rigorously evaluated, it was impossible to determine with any degree of certainty the extent to which participation in the program generally, or in a certain program activity specifically, contributed to the enhancement of these qualities. The only practical alternative was to examine the results of participation in the program as a whole versus nonparticipation, and this opened up another methodological problem—the lack of control or equivalent comparison groups.

As we said earlier, most large-scale social action programs defy rigorous experiment, and EOPS is no exception. It is impossible to deny the program to some people in order to form a control group. One does not assign people to treatment and nontreatment groups where financial aid is concerned, and it is too difficult to develop a placebo program that is different from the program being evaluated and yet of equal benefit to the participants. EOPS is a conglomerate of individual programs, and each needed to be evaluated separately. An experimental design would have required a control group of nonparticipants for each local project.

It was obvious that a classical experimental design could not be implemented. It was equally impossible to identify an equivalent group for comparison purposes since, ostensibly, persons most in need of financial and academic assistance were those recruited to the program. To identify a group of disadvantaged students who were not enrolled in the program would have been equivalent to admitting that they were not as needy.

---

[11] As it turned out, even the accountability phase was not simple, since many people disagreed about the qualifications of the target population and what constituted a "disadvantaged" person.

As more and more problems and issues emerged, the requirements for the design became more complex. The diversity of settings in which the California Community Colleges operate had led to tremendous variations in program orientation, style, and implementation. Different colleges had adopted different approaches and somewhat different emphases in programming in response to their different needs and goals, as well as those of the community. Because of these differences, it would have been misleading to evaluate EOPS from the state level or on the basis of a few selected programs.

There were, in effect, about 95 distinct programs.* It was clear that an important contribution of the evaluation would be a description of the various program approaches and the forces that contributed to the different program shapes.

Other problems resulted from the fact that the EOPS program itself had changed over the years. The original objectives outlined for the program envisioned EOPS as a special, separate entity with a full array of financial aid and academic and personal support services on each campus. Because of the community colleges' historic charge to be responsive to local community needs, and because of increasing federal aid programs, EOPS had evolved so that on many campuses it was no longer distinct from similar programs and services available for all students. Nor did every campus necessarily offer all of the originally intended service components or emphasize them in ways called for in the 1968 enabling legislation. In short, the goals and activities of the program, as well as the criteria for program success, had changed appreciably over the years. The programs did not exist in isolation; they were part of a community college—a functioning institutional system—and as such they were subject to the workings of the system as a whole. Changes in any one part of the system influenced changes in all of the other parts, and reciprocally, the EOPS program impinged upon the institutional environment if for no other reason than that it existed. To try to ferret out specific outcomes attributable only to the program was a Sisyphian task.

A third problem that emerged shortly after the study began was the discovery that different groups of people at different levels of the program and college hierarchy held quite different values and attitudes concerning both the nature of the program and its major purposes. Some saw the purpose of the program primarily as a means to increase the number of minority students in the postsecondary population; some saw it as a means to placate the colleges' liberal constituency. Some groups stressed quality over quantity, believing that the goal of the program was to make the greatest impact on the lives of the people participating irrespective of their numbers. Other groups believed that the program should process as many people as possible in the most economical manner. Still others saw the program's purpose as providing an education to a large group that formerly did not receive one.

---

*Another real world lesson is never to leave data on the floor. The custodian threw out all of the data that had been neatly stacked on the floor at one college and it had to be eliminated from the study. The final sample size was 93.

Not only were the criteria for program success different among these groups, but they expected to receive quite different information from the evaluation. The legislature and the Board of Governors, for example, wanted to know if EOPS students (supposedly "high risk, multiply disadvantaged") maintained grade-point averages and retention rates comparable to students who did not participate in EOPS. The statewide community college office was concerned with the coordination of the program and relationships between campus program personnel and the statewide office. EOPS directors and staff on the campuses were concerned about program delivery and wanted to know if students were satisfied with the support services. Faculty and administrators had still other concerns.

To complicate things even further, policies governing community college enrollment in California decree that anyone who has a high school diploma or is over the age of eighteen may enroll. This means that, although some records are kept once a student is enrolled (in most cases college grade-point average), even minimal entry data is unavailable for many students. Retention data are complicated by the fact that students drop out, stop out, transfer to other community or four-year colleges or obtain program certificates in lieu of Associate of Arts degrees. They may also become ineligible for EOPS any given term due to lack of credits or failure to fill out required renewal forms. Thus, rigorous documentation of educational outcomes, and particularly follow-ups of students' subsequent academic work, have not been a hallmark of the community colleges' data collection practices.

The ideal design for the EOPS evaluation would have been to determine long-range outcomes such as the extent to which the EOPS students became happy and productive citizens, a pay-off too far in the future for the program's immediate evaluation needs. In the absence of longitudinal data, it could only be assumed that present attitudes and behaviors were in some manner or another indicative of future attitudes and behaviors. In compromise, these considerations were incorporated in the survey given to the student samples.

Finally, by its nature, evaluation is a political activity. It provides information for decision makers and legitimizes their subsequent decisions. Where decision making is in itself political, involving the allocation of power, authority, position, or resources, evaluations frequently result in a reallocation of resources. In this case, although there seemed to be no question about the continued funding of the program (and shortly after the study began, and for no apparent reason, the governor increased the entire state EOPS budget by 50 percent), many of the people connected with the program at both the state and local levels were fearful that the funding was in jeopardy, that the proportions allocated to the various program components might be shifted, and that the evaluation results might seriously endanger the program.

An evaluation approach was needed that would take all of these factors into consideration—a design that would be comprehensive; attentive to both process and product; sensitive to the political nuances surrounding the program and the consequent fears of a good many people; address the different information needs of various constituencies; allow for changes in the goals; incorporate the different values, perceptions, and criteria of different groups of decision makers with varying levels of power and influence over the program; and compensate for the lack of pretest data on the participants. The design also had to be flexible enough to accommodate 93 different programs and to be implemented within the constraints of a minimal budget and a one-year time frame. At that point, the authors wrote a paper entitled "How to Evaluate a Complex, Multi-campus Program in a Large State System in the Real World of Higher Education where Campus Projects are Diverse, Political Pressures Intense, No Control Groups Can be Formed and No Evaluation Model Fits: or, Campbell and Stanley, Where are You Now?"

The Strategy: The decision was made to develop a holistic approach to the evaluation that emphasized careful documentation and description of processes and activities and at the same time focused on actual outcomes irrespective of prespecified objectives or criteria. The design guided the procedures. The evaluation of outcomes necessitated quantitative data from students, faculty, and administrators. The description of processes required that representative programs be observed as functioning units. The holistic evaluation designed to meet these information needs proceeded in two phases.

The first phase consisted of a comprehensive survey of randomly selected samples of EOPS students, administrators, faculty, counselors, program directors, members of local advisory committees, superintendents of multicampus districts, and non-EOPS students. Since a major criterion for success, as defined by the Board of Governors and the legislature, was that EOPS students perform as well as other students, the relevant comparison group for the study was the population of non-EOPS students enrolled in the colleges. In this case, a nonequivalent comparison group was not only appropriate but also essential to the purposes of the study.

The purpose of the survey was to compare the characteristics, experiences, perceptions, and attitudes of representative samples of EOPS students with those of non-EOPS students and to examine the attitudes, values, and opinions of the program held by both EOPS and other college staff members. Contrary to many surveys used in social research, however, this one was not a "fishing expedition." Rather, it consisted of very specific criterion instruments developed to measure each of the pre-established objectives set by the enabling legislation and the Board of Governors.

The second phase of the study consisted of intensive (and extensive) case studies of twelve colleges which were systematically selected to represent the diversity of the California Community Colleges in terms of size, geographic region, urban/rural setting, ethnic mix and programming emphasis.

The Survey: In order to develop relevant questionnaire items that would identify outcomes and processes, the full range of issues and questions surrounding the study were first outlined according to all of the program documents—the enabling legislation, the Board of Governors' Statements on Policy and Goals, Title V of the Education Code, volumes of documents, data applications for funding, and previous in-house evaluations of programs provided by the Chancellor's Office. These documents not only helped enumerate key issues and questions, but they also provided a historical perspective of the development of the programs on the different campuses and identified the forces that shaped their implementation and subsequent maturation. During the period of time in which the instruments were developed, frequent meetings were held with the Chancellor's Office staff and selected EOPS directors in order to more fully understand the attitudes and behaviors of key groups involved in the program.

One hundred forty-six questions were cast into questionnaire items appropriate to each sample. The design called for comparisons of the different samples, and consequently, there was much overlap between instruments, particularly with respect to attitudes and opinions regarding the program in general and the campus situation in particular. The preliminary set of questionnaires were pretested, and lengthy discussions were held with representatives of each sample group who suggested additional items, revised items, and deleted still others. The students were especially helpful in identifying words that had hidden or slang meanings and otherwise clarifying the language of the items for the student population. In the process, they eliminated the unintentional but, nevertheless, insidious "educationese." The revised instruments were submitted to the statewide office for review, and after the reviewers' comments had been incorporated, they were finalized and printed in booklet form, color-coded to represent the different samples.

In addition to the survey questionnaires, a Basic Data Sheet was developed for the colleges in order to gather baseline data on the vital statistics of the local college and the campus programs. This information included enrollment figures, funding allocations, staffing and, where available, data on students' high school and college grade-point averages and retention.

During the final period of refining and printing the instruments, the study team also began to work with representatives from each of the 93 colleges. The community college presidents had each designated a liaison person to coordinate campus activities and facilitate communications between the evaluators and the campuses. In some cases, the appointed liaison was the campus EOPS director; in others, the Dean of Student Personnel Services; and in a few cases, a faculty member served as a liaison.

Six regional training workshops were conducted by the evaluators in order to acquaint the liaisons with the purposes of the evaluation, the design, the purposes of the instrument and, on a practical level, the procedures they were to use for selecting local samples and administering the questionnaires. There were several pay-offs from these workshops. In addition to giving the liaisons specific instructions about administering the surveys, the workshops provided a valuable opportunity for the evaluators to meet the people from the campuses, answer their questions, and secure their trust and cooperation. In turn, their cooperation helped gain the interest and involvement of a broad cross section of community college personnel, and as a result, although the survey instruments were of necessity quite lengthy (ranging from 14–20 pages), response rates for all constituent groups were phenomenal—ranging from 70 to 90 percent. All of the Basic Data Forms were also completed correctly, a feat not easily accomplished.

The Case Studies: At the heart of social programs is really the issue of institutional change and the degree to which efforts at change succeed or fail. Quantitative data alone cannot adequately determine the extent to which any particular institution brings about change. In order to obtain this type of information, the second phase of the holistic evaluation strategy consisted of a series of site visits to 12 case-study colleges. Twenty-five colleges were first nominated by the statewide office to represent diversity in terms of size, region, number of colleges in the district, type of district (that is, single or multi-campus), ethnic composition, average family income, and ethnic composition of the surrounding community and scope and emphasis of EOPS. From this list, 12 colleges were selected as case-study sites, and all accepted the invitation to participate. Preliminary visits were made to meet each campus liaison, arrange for lodging, and clarify logistical arrangements for the site-visit teams.

In keeping with the goal of involving different constituencies and persons at different levels of decision making, nominations for site-visit team members were solicited from some 420 community college personnel, including superintendents/presidents, deans, faculty, vice presidents of student services, heads of counseling, EOPS directors, and officers of the state EOPS student organization.

In all, 497 persons were nominated to fill the 30 team positions—six teams of five persons each in addition to a member of the evaluation staff. Persons who received three nominations or more were invited to indicate their willingness to serve on the team, and 115 persons accepted. A final group of 30 persons was selected so that each team included a president, a dean-level representative of student services, an EOPS director, a member of the faculty, and a current or former EOPS student. Women and minority persons were represented on each team and, with only two exceptions, team members were assigned to site-visit campuses outside of their home regions.

There were actually several purposes of the site visits. First, as charged by the Board of Governors, a major purpose was to describe the ways in which each college implemented the activities and services designed to achieve the objectives specified in the initial legislation. A second, related purpose was to document how effective each college had been in achieving those objectives. In order to provide the information necessary for program improvement, it was also necessary to investigate the structural and staffing arrangements of the program and program features and characteristics of the college and community that appeared to be related to program effectiveness, and to determine the functional relationship of EOPS to other programs within the institution. Finally, an important purpose of the site visits was to determine the extent to which data gathered in the surveys accurately reflected conditions as observed by the site-visit teams and reported by the different persons interviewed.

Holistic evaluation demands a delicately balanced investigation. Relying too heavily either on a set of outcomes or the perceptions and opinions of different groups may give a wholly unrealistic impression of the actual program operation. How the staff and participants feel about a particular program in which they are involved matters a great deal. Are the services suited to their needs? Are they treated differently in other areas of the institution because of their participation? What are the physical arrangements for the program? What factors seem to be most related to participant satisfaction with the program? It is simply not possible to disentangle completely the attributes of the process and the quality of the outcomes that they generate. Program evaluation must include an understanding of the particular program in the local sense, and such an understanding can only be gained by on-site experience and systematic observation.

For example, an important finding that resulted from examining on-site structural and staffing arrangements was that on some campuses the EOPS offices were cramped, dismal, unattractive holes-in-the-wall located at obscure corners of the campus far away from either the central administration building or the social gathering area for students. Both EOPS staff and students reacted verbally to this "second class" treatment. The condition and location of the campus EOPS office, moreover, was consistently related to the college's commitment to EOPS as well as its perceived value to the faculty and staff, and this, in turn, was strongly related to students' satisfaction with their college experiences generally and their experiences with EOPS specifically. In other cases, the style and orientation of the EOPS director was related to both the direction of the program and students' attitudes.

In fact, information gained during the site visits demonstrated that the cluster of variables that came to be called a college's "emotional" commitment to EOPS was often more important than its financial commitment in affecting students' feelings of satisfaction and their social integration into the college—one of the major objectives of the program. If these elements had been omitted from the study, a valuable source of information, which in many cases explained differences in outcomes and pin-pointed areas needing improvement, would have been lost.

Each team visited two colleges, spending two and one-half days on each campus. Each visit was immediately preceded by an eight-hour orientation meeting during which the evaluators clarified the purposes of the site visits, the methodology and rationale for the interview schedule, and the general procedure to be followed during the site visits. Each team member was given a detailed outline of tasks and a set of questions to be investigated for each task. Formal sessions were conducted with presidents, a cross section of other administrators, faculty members, counselors, current and former EOPS students, members of governing boards, and representatives of local advisory committees, community schools and agencies. In the case of multicampus districts, a top-level representative of the district office was also interviewed. At least two team members participated in every interview session in order to assure inter-rater reliability. In addition to the formal sessions, site teams observed the EOPS staff in action, chatted informally with students and staff at the tutoring and counseling centers, and generally observed the overall campus environment.

The team members met for long sessions each evening to review and integrate their notes. A statement of major observations was presented to officials of each college prior to the team's departure from the campus. When all the site visits were completed, team members each drafted a profile of the college's EOPS incorporating their own opinions as well as information obtained from the interviews. Drafts were then compared and composite profiles developed by the evaluation staff. Unlike most evaluation reports, particularly those derived from experimental designs, data gathered in the two phases of a holistic model are presented necessarily in a two-volume report, with the first volume consisting mainly of analyses and interpretations of quantitative data and the second volume containing the narrative case-study profiles. A major weakness in holistic evaluations arises, however, when the process data and outcome data yield contradictory information. This is a particularly difficult problem to resolve when the balance between the two forms of data has been conscientiously maintained, and the evaluators can only rely on their intuition as to which data are more likely reflective of the "true" situation. The only solution, of course, is to present both sets of data, acknowledge their differences and withhold judgment unless a strong case can be made for the superiority of one set of data over another.

In the EOPS evaluation, the confluence of findings between the survey and case-study data was amazingly high. As a result, some information gathered at the site visits was also integrated into the first volume where it corroborated data gathered from the surveys or directly from the colleges. In the few cases where the data were contradictory, both sets of information were presented and their sources identified.

Summary: The strategy of involving a large number of people from the beginning of the evaluation and of consulting with representatives from key groups at different levels of influence and responsibility permitted a wide range of criteria for measuring program effectiveness to be included in the study and guaranteed that the evaluation was both site-specific at the local level and yet met the requirements of decision makers at the state level. An important offshoot of the "people-involvement" process was that it served to reduce, and in most cases eliminate, people's fear of the evaluation and the outside evaluators.

The fact that intensive site visits were made and case study descriptions prepared assuaged the concerns of "process" people who were initially suspicious of the evaluation. The quantitative and survey data gathered within the context of the original program objectives garnered the support of the outcomes-oriented cohort. As a result, the level of cooperation from all groups was impressive.

Finally, and perhaps most importantly, the combination of case study and objective-based data and the widespread participation of people in the study as liaisons, consultants, Advisory Board members, and site-visit team members had a significant effect on the use that has been made of the findings and recommendations. When evaluation is part of a process of planned change, the utilization of the findings in decision making is a key concern. And when recommendations are based on multiple indicators gathered from a wide variety of sources, there is little doubt as to their veracity and little resistance to their implementation. The EOPS evaluation report never gathered a speck of dust. The statewide office moved to implement many of the recommendations less than a month after the report was completed, and several campus staffs began making changes based on suggestions made during the site visits even before the study was completed.

There are many reasons why evaluation results are seldom used. Rarely does an evaluation study come up with a revolutionary and unequivocal set of findings that can be used to pinpoint exactly the areas needing change, define what kind of change is needed, and estimate with complete accuracy the true worth of the program for all participants. More often than not, evaluations yield findings that can be interpreted to mean that in some circumstances, certain kinds of programs may be effective to some extent with some kinds of people. Far from being definitive and unequivocal, the findings are more often tentative, ambiguous, and site and time specific. Weiss (98) suggests the following three conditions as contributing to the lack of utilization of evaluation findings: 1) the results do not match the information needs of the decision makers; 2) the results are not relevant to the level of decision maker who receives them; and 3) the results are ambiguous, and a clear direction for future programming is lacking. We suggest that still another reason for the infrequent use of evaluation findings may be that the recommendations and/or suggested directions are too massive—akin to metamorphic change *a la* the CIPP model.

While it is still too early to tell what changes will be brought about by legislative action as a result of the EOPS study, the fact that the statewide office has already begun implementing several of the recommendations made in the report attests to both the genuine concern on their part for program improvements and also to the fact that the changes suggested were reasonable and practical.

Project Head Start: A Case of What Went Wrong

Head Start is a large-scale, broad-aim, federally funded social-action program in which a variety of services (instructional, medical, dental, psychological, and nutritional) are provided for poor preschool children. Head Start began in the summer of 1965. Like EOPS, it grew rapidly, and by 1967 approximately two million children, the majority of whom were from minority backgrounds, had participated in the program.

Also like EOPS, the nature of the program and its goals posed many difficult problems for evaluation. Since the program seeks to bring about major political and social changes, its evaluation cannot be approached as if it were a traditional program designed to bring about traditional, incremental educational change. The goal is broad; the program is directed at millions of children all over the country; program delivery varies greatly from community to community; the program was not created locally, but by the federal government; and the amount of money invested is enormous. Unlike EOPS, however, evaluation was planned for from the beginning, and several evaluations were carried out by the program's Office of Research and Evaluation and its 13 Evaluation and Research Centers in universities throughout the country. Still, from the beginning, evaluation met stumbling blocks. Most studies were local or regional, and it was impossible to determine the extent of the program's overall effect or even the effectiveness of the different types of local programs.

This case study concerns the national evaluation of Head Start conducted for the Office of Economic Opportunity by Westinghouse and Ohio University (14, 104). The study included a national sample, comparison groups of nonparticipants, multiple measures of cognitive and affective development, and an evaluation of program outcomes through the third grade. The purpose of the evaluation was to make an overall analysis of the program, providing information for policy makers to decide if the program should be continued, modified, or if parts of it should be dropped. The evaluation did not include investigating the effectiveness of local implementation procedures or the delivery of program components.

The basic question that the study addressed was: Do children in the first, second, or third grade who have had Head Start experience, either summer or full year, differ significantly in their cognitive and affective development from comparable children in those grades who did not participate?

Sample: A national sample of 225 program sites was randomly selected for study from the 12,927 Head Start Centers in operation during the 1966–67 school year. Only 104 centers were ultimately confirmed as investigation sites due to the absence of appropriate control groups, lack of staff during the summer phase of the program at some sites, and the fact that some of the programs had been in operation for only one year. Other centers were excluded because some of the schools in their target areas declined to participate in the study.

Procedures: Fifty-five interviewers were recruited and given a one-week training course to prepare them for the field studies. They were each assigned to two sites spending approximately three and one-half weeks at each center, meeting with three groups of people during their visits: local Head Start officials, school administrators, and parents of both Head Start and control-group children. The progression of their activity at each site was as follows:

1. Interview the Head Start official.
2. Obtain a master list of pupils who had attended the center in the specified program and year.
3. Visit the local schools and identify all Head Start children still enrolled.
4. Draw a random sample at each of the grades represented.
5. Consult with Head Start and school officials.
6. Study all available records to identify a control population, matching each Head Start subject with a control subject on the basis of sex, race, and kindergarten attendance.
7. Interview the parents or guardians of each Head Start and control-group child.

8. Arrange for the testing of pupils to be conducted subsequently by field examiners.

9. Write a field report and complete a questionnaire on field experiences.

The Head Start officials, school administrators, and parents were all very cooperative. The only problems encountered with centers arose in those cases where poor records had been kept; only 10 school systems were considered uncooperative, although the cooperation of others reportedly required exceptional diplomacy on the part of the field interviewers. Over 90 percent of the parents were reported to have been "very cooperative" or "cooperative." The most serious problem faced in the study was finding parents who had moved or been relocated by urban renewal projects. In Appalachia, one field worker, mistakenly identified by a parent as a "revenue" agent, was shot at! But other than this somewhat humorous incident (at least in retrospect), surprisingly, everything went according to schedule, and overall resistance to the evaluation was considerably less than the investigators had anticipated.

Results: Briefly, the major findings of the study were that the summer Head Start programs were not effective, and the full-year programs were marginally effective. The major recommendations were therefore obvious: The summer program should be phased out, and the full-year program should be continued and improved—very simple and very straightforward. But, as many readers are aware, this evaluation and its findings became the subject of a heated controversy that swept the country, damaging the public's faith in national evaluations, and the residual effects remain to this day.

The fires of the controversy were lit when the findings of the study, presented as the first draft of the final report to the Office of Economic Opportunity for review and comment, were released to the public prematurely. The findings were in preliminary form and excluded several statistical analyses that were subsequently added to the final report. To make matters worse, these preliminary and incomplete findings were reported as definitive to Head Start schools, officials, and concerned parents by the news media, not the evaluators.

The attention focused on the study fanned the fires and served as a rallying point for proponents of the program, who gained additional media time and space to critique the study. The essentially negative findings of the evaluation provoked local testimonials in defense of the projects, as well as newspaper editorials and other reactions from those in the "early intervention" philosophical camp. The study was scrutinized and attacked as no study had ever been up to that time. Scholarly journals burdgeoned, and conferences overflowed with critiques of the methodology, statistical procedures, and outcome criteria selected. Most of these criticisms dealt with the defects inherent in studies of social-action programs (11, 16, 50, 80, 106)—criticisms that could be applied to many evaluations. In fact, McDill, McDill, and Sprehe (53) question whether such strong criticism would have been forthcoming from so many quarters if the evaluation had been more favorable.

Summary: This case study was not used merely to provide an example of the failure of experimental design, but rather as a contrast to the previous case study of the EOPS evaluation and to illustrate what can happen in the case of a real world evaluation that is potentially political and emotionally volatile, and what pitfalls can be avoided. Cicerelli (15), too, has reflected upon the study and suggests that if the principles of transactional evaluation had been applied, much of the misunderstanding and conflict that ultimately defeated the evaluation could have been greatly reduced, if not avoided altogether. For example, the Head Start evaluation clearly threatened the jobs of many people, and this issue should have been confronted. Although the direct participation of all parties concerned would have been impossible (if not logistically, at least economically), much more contact and "checking in" with representatives of the various constituencies could have been carried out throughout the evaluation. At the same time, although reaching consensus regarding the criteria by which to judge the effectiveness of a national program is equally impossible, greater efforts at including a broad array of criteria, including some that were acceptable to each constituency, could have been made.

Here again was a case of differing values and perceptions and the absolute necessity of clarifying these differences prior to the evaluation. The government and the evaluators agreed that the cognitive and affective aspects of the Head Start childrens' development were the most important objectives of the program. But others felt that the voluntary parental involvement and the nutritional benefits gained by the children were equally important, especially in the case of the summer programs. In fact, the same criteria were used to evaluate the two separate program components, summer and full-year, although both the objectives and the length of time available to work toward their attainment were different for the two components. The centers weren't consulted regarding the criteria by which the program was to be evaluated, and yet, the local programs varied depending upon what their center's primary, secondary, and short and long-term objectives were for the program, particularly during its developmental stages. Many centers may have been directing more attention and energy to other objectives in response to local needs, not necessarily to the exclusion of, but in addition to the objectives defined for the overall program. Since they were not consulted regarding their objectives and program emphases, valuable information was missing from the evaluation.

Field interviewers were assigned to sites on the basis of their complementary ethnic or racial backgrounds and/or their multi-linguistic abilities. In addition, however, local people could have been effectively involved as research coordinators to assist in the field work, just as local liaisons were used in the EOPS study. As it was, communications between the center and school staffs and the investigators were poor. Only the top administrators of the centers and the schools knew about the study before the field interviewers appeared on the scene; teachers, counselors, and parents were almost totally excluded by design, if not intent. Local persons would not only have been able to deal more effectively with the resistance to the study on the part of local school staffs and improve channels of communication, but they also could have been used to share some of the preliminary findings of the investigation with local Head Start and school personnel and parents, perhaps helping to avoid the furor created by the prematurely released report.

Following the transactional model, protagonists and antagonists should have been brought together both in the planning stages of the evaluation and during its implementation. By getting these groups, or at least representatives of them, involved from the beginning, their resistance to the evaluation could have been stemmed. The purpose of the evaluation was to provide information to decision makers regarding the future of the program. Obviously, a decision to eliminate or greatly reduce a program such as Head Start would be threatening not only to the target population but also to the staff employed in the program. Transactional evaluation principles could have been used to reduce the consequences of this threat by recognizing it, bringing it to the surface, and enabling the different groups to confront their conflicts and resolve them.

The unfortunate release of the draft report probably could not have been avoided by any methodology or clever technique, but its impact would have been reduced if representatives of the different constituencies had been involved and if local persons had been participating in the evaluation. The findings and recommendations might have remained the same, but the different views would have been acknowledged, the sources of information would have been apparent, and the intents and purposes of the evaluation would have been clear.

Curriculum Evaluation

Schools are particularly neglectful of curriculum evaluation. This may be due, at least in part, to the problem of defining what a curriculum is. According to Stake (85), "a curriculum is an educational program." An educational program is fairly easy to identify in the public schools where one can define curriculum as an integrated system of learning materials, activities, and experiences. However, as Dressel (21) points out, "in higher education, the meaning of curriculum is far less explicit." When the term is used in a postsecondary setting, it can be referring to all courses offered in a particular institution, to those contained within a particular department or field, or even to an individual student's course of study.

There are also many different ways in which a curriculum can be structured. It can be based on an assembly of courses that are deemed necessary to meet certain job requirements; it can be formed from the basics of a particular discipline or the specialized interests of the faculty in a department; it can be designed to meet the needs of a professional or technical program; or it can be developed as the result of a systematic specification of outcomes (21). But, regardless of the way in which a curriculum is developed, it must be updated and revised. It must therefore be periodically evaluated.

Unfortunately, curricular change is seldom followed by rigorous evaluation to determine its effectiveness; even more rarely is it preceded by a systematic assessment of the actual need for changes or the directions they might take. Responses to curriculum evaluation often take the form of either cosmetic changes or defenses of the *status quo,* or both, since most evaluations are designed to view the curriculum only in its own light without regard for long-range school or program goals. Even where specific goals have been defined, curriculum evaluations should not be based merely on their attainment. The goals themselves must be evaluated in order to determine their worth, relevance, and interrelationships within the context of both the overall program and the system.

There are several problems that typically mitigate against systematic curriculum evaluation. First, many faculty members view curriculum evaluation as an imposition on their inalienable rights as teachers. In particular, if the curriculum is based on their specialized interests, they view its content and substance as sacrosanct. Evaluation implies judgment, and many faculty are threatened by a process that may well point out deficiencies in programming or areas in need of improvement for which they are responsible or in which they are involved. If the results are negative or suggest changes with which faculty do not agree, they will often simply not accept the results, finding fault either with the evaluation or the people who conducted it.

Finally, although the motives for evaluation should always be scrutinized, they are particularly important in the case of curriculum and instructional evaluation. If the motives or reasons for the evaluation are not explained and accepted by faculty, they may feel that there is some potentially harmful outcome to be avoided; view the evaluation as "busy work" and not take it seriously; or view the evaluation as "management ordered," and refuse to cooperate. Resistance to change may be a contributor to any of these problems, which may in turn be used by faculty as acceptable excuses for their resistance.

Most, if not all, of these problems can be overcome and faculty can become an important part of the evaluation process, assisting in the planning, implementation, and analysis of results. The key is to involve faculty from the beginning, discussing with them the reasons for the evaluation and the potential payoffs from it and giving them time to become comfortable with the persons who will be directing the evaluation. The following case studies illustrate the problems and the successes of various forms of curriculum evaluation.

An Evaluation of a Professional School Curriculum

Overview: A new assistant dean for academic affairs was appointed at the school of dentistry, and his major concern was the curriculum. Although the curriculum had changed over the ten-year period since the school was established, the rapid growth in courses, students, and faculty had precluded rigorous assessment of its effectiveness. The new dean had had previous teaching and administrative experience at two of the most innovative dental schools in the country, where evaluation was the basic ingredient of educational improvement, and the school looked to him to direct the much needed curricular revision.

The dean, in turn, contacted the authors to explore ways by which they might assist him. At their first meeting, four weeks before the fall quarter began, he clarified his intent. He wanted to know how effective the present curricular structure and its offerings were in accomplishing the goals of the school, meeting the needs of the students, and most important, preparing the students to be practicing dentists. Our task was to draw up a plan for the evaluation and present it to him in two weeks. If the plan was accepted, the project would begin as soon as the school year started.

During the next two weeks, we examined the school's extant goals, reviewed accreditation reports, and interviewed small, representative samples of both faculty and students. At the second meeting, an outline of the evaluation strategy and objectives for the project were presented to the dean as follows:

1. To systematically develop measurable curricular goals for the school and departments based on graduate outcomes;
2. To evaluate the attainment and relevance of these goals on the basis of actual graduate behavior and attitudes in their practices;
3. To make appropriate changes in the curriculum based upon the information gained from both the study of graduates and the goal formulation process itself;
4. To assist faculty in planning, developing, and evaluating instructional strategies relevant to the curricular and instructional goals; and
5. To establish an on-going evaluation program to facilitate a continuous process of curricular change and renewal.

The initial analysis revealed that the school goals, as stated, could contribute little, if anything, to an evaluation of the curricular program. Like most so-called educational goals, they fell into two types of statements: "the school will provide . . ." and "the graduates will be good dentists" (or the equivalent). The former type of goal is met simply by providing whatever is to be provided, and evaluation merely consists of a double check of that provision. The latter type of goal is so global and vague that it is impossible to measure its attainment; evaluation is equally impossible. Neither type of goal is reflective of or dependent upon curricular practices. Clearly, the first priority for the school was to establish goals that were specific, measurable, and directly related to the curricular program. The objectives for the evaluation were accepted by the dean, and that is when process became a high priority as the foundation for the project.

The best laid evaluation plans can come to naught if the support of the people involved is lacking. To be effective, curriculum evaluation in particular must be conducted and perceived as a cooperative, collaborative venture, not as an activity imposed upon the majority by a select group of individuals. In this case, the faculty had had many negative experiences with evaluation "experts" over the years, and there was little reason to assume that they would cooperate. Their cooperation had to be earned.

The support of one very important person was obtained, but he did not involve himself in the evaluation in any way other than approving the funds necessary to conduct it. This person was the dean. He felt that the curriculum of the school was taught *by* the faculty *for* the students, and that they—the faculty and students—should together analyze it and recommend changes within a supportive, but neutral environment. He introduced us to the entire faculty at the first fall faculty meeting, reiterated his complete support of the project, and did not ask for, or receive, any further communication from us during the entire first year.

The Chronology of the First Year: In order to establish the process for developing the school and departmental goals, as well as mechanisms for evaluating their attainment, it was agreed that an existing faculty committee would work with the evaluators rather than creating a new, additional structure for the project. The standing curriculum committee appointed a subcommittee composed of an administrator, two faculty members, and one student.

In order to provide a framework for the evaluation, the following assumptions were defined at the first working meeting:

1. The goal of curricular renewal is the improvement of teaching and learning.
2. Any really meaningful changes in the curriculum and, ultimately, improvement in the teaching-learning process, must be fully integrated with a rigorous, comprehensive evaluation strategy.

3. The focus of evaluation must be on outcomes—in terms of student achievement and satisfaction; faculty motivation, development, and satisfaction; responsiveness of course offerings and curricular sequencing; and, finally, outcomes in terms of the total school environment.

At that same meeting, the subcommittee reviewed and ratified the objectives of the project and agreed upon the procedures that would be used to accomplish them. The first step was to solicit ideas for first-order school goals through interviews with the faculty and students, and on the basis of these conversations, each committee member would generate a list of tentative goals for consideration.

Although the components that comprise a measurable objective were reviewed at the meeting, the committee members returned for the subsequent meeting two weeks later with a lengthy hodgepodge of vague ideas and global, motherhood-type statements similar to the vacuous descriptions of most school catalogues. Several hours were spent distilling their essence, collapsing them, and rewording them into goals that were at least semi-measurable and based upon graduate outcomes. During the following week, the goals were further refined and presented to the committee for review. Changes in wording were explained, and approval was obtained for each change. When all of the goals were in acceptable form and accurately conveyed the intents of their "authors," it was agreed that they should be circulated among the faculty and students to gain their reactions and acceptance.

The tentative goals were sent to every full-time faculty member and 25 percent random samples of the student body, each drawn representatively from all four class levels. Everyone was asked to review each goal and suggest criteria that they would accept as evidence of its achievement. Response rates were 90 percent from the faculty and 80 percent from the students. Consensus on the goals ranged from 75 to 95 percent for the faculty and from 80 to 95 percent for the students, and many suggestions for measurement criteria were obtained. Some of the respondents also suggested rewording goals they agreed with in essence, and some suggested additional goals for consideration. The tabulated results, along with the suggested word changes and lists of criterion measures, were circulated again to get another reading on the goals and a first reading on the criteria. This time, part-time faculty and faculty who held joint appointments in other schools were also included.

Again, responses and consensus were overwhelming, both for the goals and the measurement criteria. The criteria were further refined and sent out once again. Then, only those objectives and measurements which received over 75 percent agreement were adopted as first-order goals of the school—the cut-off point previously agreed upon with the committee and the faculty.

Everyone agreed that acceptance of the goals by at least three-quarters of the faculty would minimize the possibility that a vocal minority would prevent their attainment.

The same process of goal formulation was then instituted for each of the school's sections (units equivalent to departments), except in this case, section representatives formed the working committees, and each section defined its own cut-off point for goal adoption. A few were lower than 75 percent, but most were higher. All of the faculty in each section were involved in the process, and measurable objectives and criteria were established for each section that were congruent with and supported the goals established for the school at large.

The first year of the project was thus completed. The school and each of its sections had a set of objectives and criterion measures to assess their attainment, and a process for curriculum development and evaluation had been established. At many institutions this would have been a one-month project. Why had it taken so long in this case?

In Retrospect: Completing the initial stage of the project took the better part of a year, but we firmly believe that the slow movement through this phase was essential for several reasons. Many of the faculty were far from receptive to the project from the beginning. They had experienced too many simplistic workshop overviews of objectives. They had been required to write "behavioral" objectives for their courses, but few faculty saw the relationship between the objectives and their teaching. Once written, the objectives were filed away only to be brought out for periodic accreditation visits. Since few faculty actually used their course objectives, the relationship between school and section goals and the curriculum was very remote.

A second reason for the slow progress was that some of the faculty who favored the project from the beginning were supportive for the wrong reasons. Anticipating minimal cooperation from other faculty, they saw this project as a way to railroad pet goals into the curriculum and thus obtain more curricular hours for their section. In the early stages of the project, it was evident that for some people an important measure of professional worth lay in the number of curricular hours for which they were responsible. A recurrent theme in introductory conversations was: "Hi. My name is Dr. So and So (no one ever had first names); my section has sixty-three curriculum hours. The national average is forty-two, you know."

In order to counteract this attitude and yet gain the faculty's cooperation, we spent the first two months of the project doing little more than visiting faculty in their offices, chatting with them in the halls, and having coffee and lunch with them in order to get to know them, explain the purposes of the project, answer their questions, and slowly gain their support. The time was not ill spent, in spite of the fact that one of us gained ten pounds and the other became allergic to coffee. Many faculty simply needed to get to know the evaluators (and in some cases, judge them) on a personal level first and as evaluators second. Others needed to get their "air time" to present gripes about the school and/or to demonstrate their own expertise as educators/evaluators. Oddly, the fact that we knew nothing about dentistry was never raised. But through these formal and informal visits, the purposes of the project were conveyed to the faculty, and they began to accept the fact that there was no hidden agenda, that we could be trusted and that there would indeed be an evaluation that *they* would help design and conduct. The formal process of goal setting and review could then begin. But soon after, the project ran into its second slowdown.

As the faculty became convinced that they really would be responsible for establishing the curricular direction of the school, they developed an almost insatiable thirst for information: How can I be sure that the objectives for my section will be good? How do we know our tests are fair? How do we evaluate clinical performance? How can I be sure that my instructional materials and methods are adequate? In response to these requests, and with the support of the still-invisible dean, we conducted a series of seminars and workshops ranging from methods-type classes on instructional techniques and student learning styles to workshops on test construction and clinical evaluation.

As a result, by the end of the first year, in addition to the goals and criteria that were established for the school and sections, a learning environment had been created in which teaching received major attention. None of this would have happened if the faculty had not been developing their own goals with an eye to attaining them through teaching.

The goals certainly could have been stated better if they had been written by, or purchased from, professional educators. And they would have been written in much less time and probably for much less money. But they might have ended up on the proverbial shelf along with the other goals and objectives that had been lying there for years. The purity of measurable objectives had been violated. But, while classically imperfect, their intents were adequately conveyed and faculty were committed to working toward their attainment. Garnering faculty support was far more important for the future of the project than producing classically perfect objectives. There was ample opportunity to reword the objectives later; there was only one opportunity to gain the faculty's trust. So the first year ended—far behind schedule, but way ahead in support.

The Second Year: The next phase of the project was inaugurated by having each section present their objectives to the rest of the faculty, demonstrating how they contributed to the overall school goals and complemented and expanded upon those of the other sections. Many overlaps and gaps were identified, and *ad hoc* joint section committees were set up to investigate and explore solutions.

Each section had been asked to send one representative to these information-sharing sessions. If all had complied, that would have meant an attendance of 23. However, so many people were interested, an average of seventy came to each of the first three meetings. Also, a 16-hour course on criterion-referenced measurement was offered and 27 faculty attended. A series of mini-courses on "What We Wrote As Objectives But Never Learned Ourselves" was introduced, and 40 people came to the first session. Since there were only about 60 full-time faculty (and 140 part-time), this represented an amazing show of support for in-service training. In addition, complaints were voiced by faculty who had classes or laboratory sessions which conflicted with the hours of the workshops and seminars. As a result, and at the request of the faculty, the dean designated one-half day each week as "Faculty Development Day." Classrooms and laboratories were closed and the school turned into an instructional laboratory. More teaching improvement classes were introduced, as well as several discipline-oriented continuing education courses. Faculty attendance, which was always voluntary, hovered around 90 percent.

And where was the curriculum evaluation that had started all of this activity? Actually, it was all over the place. After hearing about a particular instructional principle called appropriate practice at one of the classes, one professor cancelled a lecture and took his students to an empty laboratory to practice. Another took his students out of a laboratory where they were practicing something he decided wasn't all that important. Still another faculty member decided that his age-old practice of giving a quiz at 8:03 every morning, for no other purpose than monitoring attendance, could be dispensed with.* Faculty were reexamining their objectives, not necessarily with an eye to changing them, but to understanding their full implications for the classroom.

Faculty were talking to each other about their teaching and how their students were progressing toward particular objectives; curricular hours were seldom mentioned. Moreover, somehow, coincidentally, first names surfaced, and the unfriendly atmosphere seemed to disappear. Many of the faculty began to work with us on a number of special projects. Some developed self-instructional materials that contained outrageously irreverent cartoons and humor (which the students loved—and learned from); others helped us prepare an objectives-based questionnaire for the graduates. Change was taking place, albeit somewhat less systematically and more serendipitously than had been planned. Evaluation and change had become a cyclical process. Evaluation served as the incentive for change; change, in turn, necessitated evaluation.

The Survey: The foundation of the formal evaluation was the graduate questionnaire, and again, everyone was involved in its development. The 40-page compendium of objectives was analyzed, and items were developed for each objective specified for the school and the sections. A draft was approved by the committee and distributed to the faculty and student body for their review. This was not a typical alumni questionnaire full of "Where are you, what are you doing, and what did you think of your education?" It was entirely dependent upon the specified measurable objectives. Because of its length, however, the faculty and students were asked to designate items that they considered "absolutely essential" so that two versions of the questionnaire, with these as common items, could be devised, thus keeping the length manageable for any one respondent. Each questionnaire still ended up being 17 pages long. They were sent to every person who had graduated the previous year, and over 60 percent were initially returned. A follow-up letter increased the response rate to 85%.

---

*The use of the masculine in all of these cases is purely reflective of the sex of each person mentioned.

We had intended to codify the data, analyze it, and prepare a written report on the findings, but the faculty were impatient. They were determined to have an immediate look at the information, so the data were simply tabulated and sent to them. At the same time, a formal analysis was prepared, but the faculty didn't seem to need it.

For the next few weeks, everyone was talking about the survey results. Special section meetings were called to discuss the implications. Requests for curricular changes were brought to appropriate committees and moved through the formal structure. Informal changes took place immediately in the classrooms. Timidly at first, faculty began to ask if their objectives were "revisable." Some were quite concerned that their most cherished objectives were being ignored by the students once they graduated. The purpose of the survey was to find out what graduates were actually doing in their dental practice. Now that the faculty had that information, they had several alternatives: they could eliminate or change their objectives; they could institute measures designed to ensure that more graduates would accept the importance of their objectives and honor them in practice. They could also do nothing. It was up to them; they had the information, and they now had the skills.

Summary: This *was* a curriculum evaluation, but it contained many more elements than are typically found in such endeavors. Some would separate the project into faculty development, instructional improvement, organizational development, evaluation, or other categories. The project may have encompassed these components, but the focus was upon one goal—the improvement of teaching and learning. In this case, these elements were companion activities necessary to accomplish the goal.

The curriculum evaluation was and continues to be a success. Student and faculty evaluations of the project were extremely favorable. They were pleased with the processes, the learning outcomes, and the results. The project has been institutionalized and is now in its fourth year. As intended, we, as external evaluators, became dispensible; the faculty continued the process of change and evaluation.

The following principles summarize what we believe are the major reasons for the project's success:

1. There was strong administrative support coupled with very low administrative visibility.
2. The evaluators were allowed flexibility in the initiation and evolution of the project (for example, the two months spent "setting the stage" and the ability to add elements to the project at the request of the faculty).
3. The evaluators were there as objective, external-change agents, removed from internal politics and with no ties to any particular constituency.
4. The project was designed to respond to the immediate concerns of faculty, giving it a credibility and an influence necessary to confront the more complex and comprehensive changes to come later.
5. A major standing committee was used to help plan and implement each stage of the project. For this reason, faculty did not feel that something foreign was being imposed on them.
6. Faculty (and students) were involved in the conduct of the project from the beginning and had basic control over its direction and outcomes.
7. Faculty were offered training in the skills required for full participation in the project. Those who wished to learn more were trained as a cadre of "in-house experts" to work with others on an individual basis and lead courses and workshops themselves.
8. The evaluation staff and budget were kept to a minimum. The two external consultants were augmented by resources already there. The dean's secretary arranged all meetings and schedules; other tasks were assumed by faculty and students.
9. Evaluation was fully integrated into all aspects of the teaching/learning/management processes of the school; it was not a mere appendage.
10. The mechanisms, processes, products, and outcomes of the project were fluid.

An Evaluation of a Public School Curriculum

Overview: A small, relatively isolated, politically conservative rural community had been transformed, because of its accessibility to a major expressway and a building boom that spanned a ten-year period, into a large suburban community with a great mix of people who held opposing views regarding educational philosophy in general and the curriculum of the schools in particular. When dissatisfaction with the schools grew to such a point that school bonds failed to pass, a group of parents and community residents, in cooperation with the school board, initiated an evaluation of the school curriculum and the district's policies, and also contracted with external consultants to conduct a separate "objective" evaluation (22).

The objectives of the external evaluation were:

1. to examine and document the competing values of various community groups;
2. to determine areas of agreement and disagreement among community residents and examine how these shaped the school program and affected school policy; and
3. to formulate plans to enhance school-community relations, reduce the level of conflict, and improve educational opportunities for all students.

Procedure: Although the term was not used, the basic principles of transactional evaluation were employed by the external evaluators, since the evaluation called for the confrontation and resolution of conflict. The first step was to identify the divergent goals, educational philosophies, and attitudes present in the community. Three activities were initiated in order to gain this information.

First, a mail survey was conducted by the citizen evaluation committee in which residents were asked to rate on a five-point scale the degree to which they agreed or disagreed with statements regarding educational philosophy, school goals, school programs and policies, physical facilities, school-citizen communication, and taxes. They were also asked to indicate what elements they were most and least satisfied with in the schools. Second, in order to determine discrepancies and congruencies between parents, teachers, school administrators and non-parent residents regarding the purpose of education, 36 members of the lay citizens' committee, 35 teachers, and 2 administrators were asked to rank 106 educational goals listed on a form commonly used throughout the country (40). Finally, formal classroom visitations were conducted in each class in the district, as well as some classes in a nearby district for comparison, by both members of the lay citizens' committee and the external evaluation team.

Responses to the mail survey were examined according to age groups, length of time residing in the community, and whether or not respondents had children in the school system. Two distinct value systems clearly emerged that were classified according to Spindler's (81) definitions of traditional and emergent values. Spindler defines traditional values as those which emphasize thrift, self-denial, postponement of satisfaction, success, and a belief that the means to it is hard work, absolute morals, and elevation of the individual as an end rather than the group. Emergent values are defined as those which emphasize sociability, sensitivity to the feelings and needs of others, a relativistic attitude toward moral norms, and a here-and-now orientation that reflects uncertainty about the future (81).

Traditional values were held by residents who had lived in the community for more than ten years before the changes in the community and the schools had taken place and by those who did not have children. People who had lived in the community for less than ten years and those who had children enrolled in the local schools held more emergent values. The two groups of residents who were at odds with each other were clearly identified. Breakdowns by age showed no significant differences. Although the beliefs of residents who had children in the schools might be considered more relevant and thus more important, all citizens are entitled to vote, and both board memberships and school bond issues are decided by and subject to the input of all. Further, pressures on administrators come from all quarters.

The results of the goal ranking demonstrated that the parent/citizen group ranked academic skills and their relationship to everyday life the highest, while teachers stressed creative, affective, and artistic goals. Some areas of agreement did emerge between these two groups; the development of self-esteem and knowledge of sociology and citizenship were important, and religion was least important for both groups.

During the site visits, parents and other citizens observed educational practices and found that many of the negative issues raised regarding the schools were grossly exaggerated, and in some cases nonexistent. The majority of teachers were in fact emphasizing rather traditional values. The consultants corroborated these perceptions. In addition, standardized tests were administered to students by the external evaluators, and it was found that the mean achievement scores were at least at grade level for all grades and considerably above grade level for the majority of grades. This very traditional measure of achievement satisfied everyone that the students were receiving a quality education, a point that illustrates the weakness inherent in a noncomparative evaluation. The primarily middle-class students were no doubt above average, and there is no way of knowing without a comparative evaluation whether these students would have scored higher had they been learning under another set of conditions.

Although some weak points were identified by the classroom observations, ideological judgments were largely replaced by data or perceptually based information. Specific practices were examined more as to their effectiveness than as to how much they conformed with value structures. As Eash *et al.* (22) summarized, "it soon became evident to all but the most hardened ideologists that the earlier assumptions were too broad, often did not correspond to the facts, and were untenable as a basis for making policy.

Ideological rhetoric was reduced, the climate for teaching and learning was enhanced, and a better relationship with a more informed community resulted."

Although the study did have some methodological limitations, seven major outcomes were attributed to the project (22):

1. Greater community interest resulted in a larger turnout for school board elections.
2. School board members were more carefully selected and candidates for the board made more effort to inform the community of their stand on specific issues.
3. Positive community interest in the schools increased as did a readiness to contribute to as well as critique school activities.
4. Citizens were better informed and demanded increased communication with school administrators and teachers, as well as a more systematic organization of school curriculum.
5. Demand increased for accountability of school administrators to both citizens and teachers regarding curriculum, student achievement, and finances.
6. Antagonism between teachers and citizens over the purpose and organization of classroom instruction was reduced.
7. A citizens' advisory committee was established to work directly with the school board and to serve as a source of input for citizen opinion.

The most obvious reason for the positive results that came from the project was that the citizens of the community were deeply involved in both the planning and implementation of the evaluation of their schools. Possibly because the external consultants were conducting a companion study, the citizens' group steadfastly attempted to make their study as valid as possible so that their recommendations to the school board would be received with equal weight. As a result, recommendations from both groups were very similar although the external evaluators called for more extensive changes than did the citizens.

Facts that were brought into the open replaced ideological rhetoric that had previously kept the two factions of residents from agreeing on school policy. But it was not just the information that helped resolve the problems in this community; it was the manner in which the data were collected, analyzed, and reported.

Needs Assessments

The process by which one identifies needs and decides upon their priority has been termed needs assessment. A need may be defined as a condition in which there is a discrepancy between the actual or observed state of affairs and a desired or acceptable state of affairs (3). In the educational world, this discrepancy can be determined by objective measurement (for example, fourth grade students are given a test to measure their skill in mathematics, and the results are compared with a set of standards expected for children in the fourth grade). The extent of the discrepancy may also be estimated subjectively (for example, a group of "judges" observe the operations of an institution or a particular program and collectively decide what the needs seem to be). In both cases, decisions concerning the desired standards and the degree of need involve value judgments.

The following case studies briefly describe two needs assessments conducted in different settings and, accordingly, using different procedures.

A Needs Assessment of a Professional School:[12] A general feeling of stagnation existed within the school, and both the faculty and administration were dissatisfied with the quality of education being provided. The faculty was also splintered, and there was no consensus regarding either the reasons for the lack of vitality or ways by which the situation could be improved. In an attempt to bring the conflicts more clearly into focus and begin to develop solutions, the authors were asked to conduct an organizational diagnosis and needs assessment.

The objectives of the project were threefold: 1) to identify critical organizational and curricular problems that directly and indirectly affected the functioning of the school and the quality of its educational program; 2) to recommend appropriate entry points for intervention strategies that would most effectively redress the problems identified; and 3) to design a program for planned change and institutional renewal that could serve as a basis for on-going evaluation and continuous improvement of the quality and effectiveness of the organization and its instructional program.

The needs assessment was conducted by a team of three external consultants who spent one week at the institutional site. As part of the analysis, they worked with faculty, administrators, and students, helping them clarify the reasons for their dissatisfaction, identify points of conflict, and explore possible strategies that would lead to their resolution.

A variety of procedures were used to gather the information necessary for the diagnosis and analysis: semi-structured interviews and informal discussions with individuals and groups of faculty, students and administrators; direct classroom observations; and document analysis. The documents provided background information regarding the history of the school and critical events in its growth; the observations and informal conversations enabled the team to explore issues in greater depth and to discover issues that had not surfaced in the interviews.

---

[12] The type of professional school in which this needs assessment took place has been withheld in accordance with the wishes for anonymity on the part of the school staff.

76

Intensive, two-hour interviews were held with approximately 75 percent of the full-time faculty in groups ranging from four to six participants. Each group was selected to be representative of different curricular areas, varying levels of faculty rank, and tenure and length of time at the school. All of the major school committees were represented in the interview groups. In addition to the meetings with faculty, special interview sessions were held with a group of students representing all class levels and with the dean and his administrative associates. An informal "drop-in" afternoon session was reserved for people who had not been involved in the scheduled interview groups, could not attend at their scheduled time, or wished to talk further.

The series of questions asked by the team was similar for all groups. No attempt was made to interpret the answers or influence the direction of the responses by delimiting the scope of the questions or channeling respondents' answers, even when they were being heard for the fortieth time. One exception was that discussion about problems related to facilities was discouraged, since a new, well-equipped building was under construction and would soon be completed.

Throughout the entire data gathering process, the focus and intent was on exploration and discovery. The goal was to find out what the school was like and how it was perceived by both staff and students. The reality of the organization as defined by its members was the primary concern, since their perceptions influenced the school's functioning and atmosphere. The procedures established at the interview sessions reflected this perspective—that is, the interviewees were the experts as far as organizational operation and functioning and the school's programs and processes were concerned. The evaluators were there to learn about the school, listening carefully to what people said and observing how they interacted with each other.

Although the make-up of the groups obviously differed greatly in many respects, there was almost complete agreement as far as descriptions of the school's operation and functioning and identification of major problems and suggestions for their resolution were concerned. This was in marked contrast to the belief that there was conflict—the assumption on which the call for a needs analysis was based.

The major problems raised by all of the people interviewed, and corroborated by the team's informal conversations, observations, and classroom visitations, centered on the pervasive lack of communication between and among the different constituencies, the fragmentation of the curriculum, the faculty's lack of training in teaching methodology, inadequacies in the testing and grading system, and the lack of administrative follow-through. This last point is particularly important.

The interviewees commented repeatedly about the lack of follow-through that characterized the *modus operandi* of the school. A workshop was presented; a project was started; ideas were generated and accepted. But no support was ever provided for implementation, and faculty who became excited about a new project or idea for change soon became disillusioned. The faculty's skepticism regarding the prospect of change carried over to the needs assessment project and, although they were cooperative—having decided to "give it one last try"—many expressed doubts that any changes would be forthcoming.

In this case, however, they were wrong, and they were rewarded for their efforts. As recommended in the needs analysis report which was circulated to all faculty, the dean approved the initiation of a long-term program of planned organizational and curricular change and evaluation. The program is designed to address the problems identified and contains those components suggested by the faculty and students who had been interviewed. The school was diagnosed to be a closed system, one that would increase in entropy and disintegration. The goal of the change program is to help the school become more of an open system by establishing a process for a cycle of self-generating change and evaluation.

The scope of the program is broad, and its chances of success will be enhanced by the continuing involvement of many people. Although it is not far removed from the dean's perception of what a change program might be, it is not *his* plan. It belongs to and will be implemented by all three of the school's constituencies—the faculty, the students, and the administration.

A Needs Assessment of a Faculty Development Program: A large community college district had established an instructional grant program that provided funds for the development of innovative approaches to teaching and instruction. Faculty in the district could write a proposal and, if selected in the competition, obtain funds to develop their design. The program had been well received, and many faculty undertook a variety of projects. Although the program had been operating for a number of years and had proven to be an excellent device for motivating faculty to examine their teaching, the director was concerned about the quality of the instructional products that were being developed. Funds for field testing were not available, so rather than implementing a formative type of evaluation, the director asked external evaluators to conduct a needs assessment of the program to determine if faculty would benefit from a special course designed to teach them the principles of instructional design, product development, and evaluation.

The first step was to review all of the project proposals that had been funded as well as all interim and final reports in order to identify the nature and objectives of each project. The investigators then combined objective measurement and case study procedures in a holistic approach to the assessment analyzing the extent of the faculty's skills in instructional design and product development. The assessment approach included a short objective test given to a random sample of faculty in each of the colleges in the district and a series of on-site interviews with every faculty member who had received a grant since the program had begun and with the local campus administrator who was supervising the program. Completed products were reviewed for content and evaluated for face validity according to the principles of instructional design. In the few cases in which student performance data were available, products were evaluated for their effectiveness in promoting student achievement and/or motivation.

The faculty's level of skill in instructional design and product development was assessed on the basis of several types of data and data sources, and discrepancies did indeed exist between their knowledge of instructional design and the quality of their products. On the basis of these findings, the investigators recommended that as a condition of receiving a grant, faculty should participate in a special program designed to teach them the basic principles of instructional design, and that such a program should be developed before the following year's grant program was initiated. It would have been unfair to discontinue the program on the basis of the needs assessment. That was not its purpose. The program had increased faculty motivation to improve instruction and had rewarded those who tried to do so, and these were specified goals of the grant program. The addition of an instructional design component in conjunction with the grant program served to increase the chances that the resulting products would be effective and of high quality.

Summary

The case study descriptions of evaluations presented in this chapter demonstrate quite clearly that there is no one cut-and-dried method for conducting them; nor is there one "best" approach to evaluation for all situations. They also provided examples of the variety of real world settings as well as the array of methodological choices available to the practitioner. Particular models were not imposed as a basis for generating evaluation questions; nor were evaluation designs picked out of a hat. Each approach necessitated a design that would address the information needs and questions required of the evaluation and appropriate to the particular program being evaluated.

Very often, the evaluation questions that need to be answered require experimental design. Is textbook A better than textbook B? Can students learn as well in condition A as in condition B? Does the program effectively accomplish our goals? Social action programs and organizational change programs, as illustrated here, are best suited for integrated types of approaches such as transactional evaluation and holistic evaluation. Evaluators must proceed in developing their designs much as a gourmet chef might go about concocting a new, delectable dish—selecting a bit here and a handful there, a dash of this and a pinch of that—combining the ingredients into a design that is suited to the particular program and its requirements and constraints.

We have talked repeatedly of selecting designs and approaches that are suitable to the purposes of the evaluation and the information needs of decision makers. But most real-world evaluations are constrained by the program setting, the budget provided for the study, and the time frame within which the evaluation must be conducted. Money is never mentioned in graduate seminars or in-service programs on evaluation. Yet, the truth of the matter is, despite the growing reverence for evaluation *qua* evaluation, few budgets allocated for it are sufficient to permit a thorough, rigorous, and comprehensive investigation. More often than not, evaluation designs are the result of compromises necessitated by limited funds and/or limited time. The case studies were selected to provide examples of these very real problems.

It should be apparent from a comparison of the case studies that holistic evaluation and transactional evaluation have many basic similarities. There are also important differences that may well become more pronounced as both approaches are refined through continued use in various settings. The common threads of holistic and transactional evaluations are: 1) persons representing key constituencies at different levels of the program and different levels of power to influence the program directly or indirectly are involved from the beginning; 2) multiple measures are used, including quantitative data and qualitative information obtained from observations and interviews; 3) there is a concern for both process and outcome beyond attainment of pre-specified objectives; 4) the study of actual outcomes is combined with naturalistic observations of what was delivered and how people interacted; 5) predetermined goals are not required nor are alternative causal possibilities eliminated in the analysis without sufficient examination; 6) experimental design can be incorporated, but where this is impossible or impractical to implement, other designs can be adapted; 7) evaluation can be viewed as either a continuing part of management or as a short term *post hoc* analysis; and 8) evaluators can serve as part of the program staff or as external evaluators outside of the program or organization. Both approaches are eclectic and flexible, and are adaptable to the needs and requirements of the particular program being evaluated and the particular information needs being addressed. They are pragmatic, common sense approaches to program evaluation that provide comprehensive information acceptable to many different constituencies and useful to many different decision makers at many levels of power.

The strategy of involving different people from the beginning of the evaluation, including some people who are antagonistic to the program or may become so, is an important part of both approaches because transactional evaluation is concerned with the resolution of conflict. In the absence of controlled experiments, the participation of program opponents increases the likelihood that biases in favor of the program will be balanced and outcomes credited to the program will be verified. Another benefit is that initial meetings with representatives of different groups serve to introduce the evaluator to a broad cross section of key decision makers. The evaluator, in turn, can use these opportunities to explain the purpose and needs of the evaluation, answer questions, involve people in the process, and try to garner their support and cooperation.

Generally speaking, neither holistic nor transactional evaluation costs more than a traditional design, and they may well cost less than large-scale experimental design. A key element in transactional evaluation, however, is that representatives of the different groups be brought together so that conflicts can be brought to the surface, confronted, and resolved. Obviously, this is feasible only in relatively small-scale studies or in large-scale studies for which budgets are sufficiently large to allow people to come together. In holistic evaluation, there is not a great deal of emphasis upon formally bringing the different groups together; whether or not it is done depends upon the particular situation. Confrontation and the resolution of conflict are not strategic parts of holistic evaluation's foundation.

The collection of data has always been valued as a respected academic pursuit. But dissemination, other than through traditional journals and scholarly association meetings, has not been a responsibility accepted by evaluators. In many evaluations, the emphasis is placed on the dissemination of information to the upper levels of management—the top decision makers only. Little feedback is provided for personnel directly involved in the program, let alone persons who are not involved directly but whose decisions nevertheless affect the program operating within their organizational jurisdiction. Transactional and holistic projects are responsive to the information needs of a broad audience—from local program staff to institutional administrators, system officials, and legislative policy makers.

Finally, holistic evaluation and transactional evaluation provide two legitimate alternatives that can be considered when experimental or quasi-experimental designs cannot be applied, and they should thus be included in every practicing evaluator's repertoire of program evaluation methodologies.

UTILIZATION, QUALITY, AND ETHICS

One further cluster of issues that must be addressed in depth concerns the use of the evaluation results. A well-designed and well-conducted evaluation improves the process of decision making; it eliminates, or at least greatly reduces, the influence of political or self-serving factors, and it provides objective, defensible evidence. Evaluation can lead to the planning of more effective programs, since data-based evidence of what is working and what is not is available to program planners. Evaluation increases the likelihood that decisions will be wise and that subsequent policy will be rational. Why, then, should the results of such a wonderful process be so universally ignored? The fact that evaluation has generally had so little impact is well documented (10, 25, 60, 66, 72, 98, 99).

Throughout this monograph, we have stressed the importance of providing information for decisions regarding program improvement and decisions regarding a program's future. But the reasons for undertaking real world evaluations are not always so rational; nor are the underlying motives always so nice. The actual use of the results of an evaluation often merely reflect the reasons the evaluation was called for in the first place.

Some evaluations are little more than public relations rituals carried out to satisfy taxpayers or other publics demanding accountability; others are initiated merely to satisfy federal or state grant requirements. These evaluations are conducted not because program staff really want to find out how well their program is working, but because they have to evaluate if they want to continue receiving the external funds necessary to continue the program. In many of these cases, program staff really don't give a hoot about the findings of the evaluation. The fact that it was conducted is enough in and of itself.

In Popham's (60) view, many educational evaluations are carried out "in a thoroughly practical milieu in which an evaluation's results will constitute additional playing cards that people will be dealing from patently political decks." Sometimes those decks are loaded. Politics is not confined to program operations; it affects both the motives for evaluation and the utilization of its results. Even the most dispassionately gathered, methodologically perfect data can be used to justify a weak program or destroy a good one. The best of evaluations can be undertaken for the worst reasons. Some are undertaken merely as a ploy to get rid of an incompetent or uncooperative administrator or staff person (60). Weiss (99) lists several other less-than-legitimate reasons that evaluations are initiated: to delay decisions; to provide support for or justify a program to "higher-ups"; to make a successful program more viable and increase the prestige of the institution; or as a means of self-glorification for the directors. Evaluations are initiated to appease program critics and because they are fashionable and lend a form of professional validity to the program (60).

Along the same lines, Suchman (91) describes the following misuses evaluation may serve: Eye-wash; White-wash; Submarine; Posture; and Postponement. Eye-wash refers to deliberately selecting for evaluation only those aspects of a program that look good on the surface in an attempt to justify a weak or bad program. White-wash refers to attempts to cover up actual program failures or errors. Submarine refers to attempts to destroy a program regardless of its effectiveness, and Posture uses evaluation only as a gesture of objectivity. Postponement is an attempt to delay needed action by pretending to seek the facts.

Evaluators should not be surprised in such cases if their reports are laid neatly to rest, albeit on a prominent shelf. And they should be amazed if their reports are not buried altogether when the results are negative or run counter to vested interests. Few administrators or program staff in the real world are readily willing to accept evaluation results that may place the survival of their program (and their jobs) in serious jeopardy. Only when organizational personnel themselves are dissatisfied with a program will they be receptive to the implications of a negative evaluation and take its results seriously. Evaluators would save themselves a great deal of anguish if they found out what the motives underlying the evaluation were and made sure that the purposes were truly legitimate before they began.

Evaluators would also be wise to be attentive to some basic procedures that seem to increase the likelihood that the results of an evaluation will be used: 1) identify potential users early in the evaluation and address issues of concern to them; 2) involve representatives from different constituencies in the process of evaluation; 3) complete the evaluation promptly, according to schedule; 4) prepare several forms of the report, including a nontechnical summary for lay audiences; 5) provide individuals whose program is being evaluated with a draft report so they will have an opportunity to critique the report and prepare a rejoinder; 6) take responsibility for presenting the findings to decision makers and interpreting them into action plans; and 7) be available for advice or assistance in implementing recommendations even after the evaluation has been completed.

The assumption in this discussion, of course, is that the evaluation report is detailed and clearly indicates specific ways by which the program might be improved. But many evaluators refuse to make suggestions or provide direction for improvement, viewing their role as one of data gatherer and analyzer only. Many evaluators provide only global recommendations that are simply too sweeping to be practical. Many evaluators make recommendations that are vague and open to varied interpretation. Yet, few of them are willing to stick around long enough to interpret their data or help translate their recommendations into action plans once their evaluation has been completed. Evaluators who abdicate responsibility for follow-through invite nonutilization of their results.

Finally, a major limitation on the use of evaluation data and a major issue that must be addressed concerns the quality of the evaluation and the evaluator. There is a tremendous gap between the rhetoric of evaluation and its demonstrated performance. According to Weiss (98), "Much evaluation is poor; more is mediocre," but, despite a few suggestions to evaluate evaluations, there has not yet been developed a formal structure or guidelines by which that can be accomplished. Guba (36) offers the following criteria for a "good" evaluation:

1. Internal validity: The evaluation information corresponds to the phenomena which it purports to describe:
2. External validity: Most evaluations are unconcerned about generalizability, but widespread application is important, particularly in the case of social action programs, and here questions of sample representativeness and the similarity of testing conditions become important.
3. Reliability: Information provided by the instruments is consistent.
4. Objectivity: Equally competent, independent judges or observers would agree with the results.
5. Relevance: The evaluation information relates to the original purpose of the evaluation.
6. Importance: The information presented in the report is important.
7. Scope: The whole story is told with a wide range of information including negative perceptions or facts.
8. Credibility: The client and other audiences of the evaluation trust the evaluator and have confidence in the sources of information.
9. Timeliness: The information is prepared in time to meet the client's needs.
10. Pervasiveness: All audiences who are entitled to it receive the evaluation information.
11. Efficiency: The cost of the evaluation in terms of time, personnel, and funds is appropriate to the utility of the evaluation information.

It is not easy to define criteria by which to evaluate evaluators. Furthermore, as the pressures for evaluation increase and more and more evaluations are required, the lure of the dollar will mount and evaluators will be faced with many ethical choices and many threats to their integrity.[13] Evaluation has become a profitable enterprise, and suddenly people from all walks of life are calling themselves evaluators in spite of the fact that they may lack training in evaluation and do not possess the technical competence to carry out quality evaluations. Having read a book or two or attended a course on evaluation does not an evaluator make. Evaluation is difficult even under the best of circumstances, and seldom do the best of circumstances occur.

Evaluation is also a high stakes game, and it is not yet a well-honed professional practice with a code of ethics or Hippocratic oath. At worst, evaluators can become whores prostituting themselves for sufficient incentives (34); at best, they can unconsciously shade information that might harm a program to which they feel morally committed. If an attorney loses a case, or a doctor a patient, the result is indeed grim for the losers and their families. But evaluations play with larger numbers and the impact of a negative evaluation is far-reaching. Programs *are* abolished and program participants *are* deprived of services that they may have felt were valuable in spite of the fact that the program was pronounced ineffective on the basis of other criteria. Program staffs lose jobs and their families suffer. Evaluators have the power to affect many lives, and their competence and integrity assumes monumental importance. Evaluators must be skilled and they must be competent. Above all, they must be ethical.

---

[13] The reader is referred to Popham (60) for a discussion of ethical issues facing evaluators.

Suppose the EOPS program really had been in jeopardy. Should such a well-intentioned program have been protected from a frugal governor even if it meant distorting some of the data? How could the anonymity promised to individual colleges have been maintained if "good" programs had to be identified in order to save them? There are an infinite number of questions such as these confronting the evaluator, and the answers are anything but simple.

In an effort to develop guidelines for educational evaluators that may be used as a code of ethics, many of the major professional associations have appointed special committees to consider the issue. A preliminary set of standards developed by the ethics committee of Division H, American Educational Research Association offers 11 statements of ethics for evaluators as follows:

1. Evaluators should be independent to the extent that they follow professional and personal standards. Evaluators should be free of political interference or coercion, limited only by general policies of the institution. Evaluators should be responsive to the needs of a client.

2. Mutual support and team work are ideal. A client-professional relationship should exist where each can have due respect for the other, but separate responsibilities. Evaluators should be accountable to the clients, but not subordinate to them.

3. Political and social contexts exist and should be duly considered when reporting findings. The true outcomes of the studies should be reported, regardless of other factors.

4. Evaluator values may be expressed in the report, but should be identified clearly as personal judgments. Values and personal biases of the evaluator should be made known to the client.

5. The evaluator has the primary responsibility for design and methodology and should make the final decisions on them. The design and methodology should be agreed upon by the user before implementation.

6. Review of the design, instrumentation, and other aspects of the evaluation by the client and fellow professionals should be sought.

7. It is an essential responsibility of the evaluator to be honest in reporting limitations and/or constraints of the evaluation.

8. Negative findings should be treated the same as positive findings when reporting to the client.

9. Release of results should be dependent upon the terms of the contract between the evaluator and client.

10. The names of individual subjects should be kept confidential at all times, in accordance with federal law.

11. The evaluator should not accept an evaluation contract when evaluator ethics and bias are at stake.

(*Division H Newsletter*, v. III, July, 1977)

A joint committee composed of representatives from AERA, the American Psychological Association, and other national organizations have developed preliminary guidelines that are currently being reviewed by prominent educational evaluators. These guidelines, which have not yet been released to the public, cover everything from the scope of the information and timeliness of the report to the fiscal responsibility, diplomacy, and formal obligation of the evaluator.

In addition to association committees on ethics, several leaders in evaluation, such as Michael Scriven, Robert Stake, and Blaine Worthen, have also begun to address the issue of evaluator ethics in their writings, and no doubt evaluators will eventually be able to turn to these documents for guidance. In the meantime, it is important to recognize that evaluation is an area that is fraught with debatable questions of ethics and moral implications. Until such time that definitive guidelines are available, evaluators must be scrupulously circumspect and conscientious. They should approach evaluation as a constructive process, viewing the goal of evaluation as improvement, and when in doubt, they should remember the immortal words of the patron saint of evaluators who said, "Let your conscience be your guide."

REFERENCES*

1. Airasian, P. W. Designing summative evaluation studies at the local level. In W. J. Popham (Ed.), *Evaluation in education: current applications*. Berkeley, Calif.: McCutchan, 1974.

---

*Items followed by an ED number (for example, ED 103 [illegible]72) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of *Resources in Education* for the address and ordering information.

2. Alkin, M. C. Evaluation theory development. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 105–117. (Originally published: *Evaluation Comment*, October 1969, 2, No. 1, 2–7.)

3. Anderson, S. B., Ball, S., Murphy, R. T. & Associates. *Encyclopedia of educational evaluation*. San Francisco: Jossey-Bass, 1975. ED 103 472

4. Aronson, S. H. & Sherwood, C. C. Researcher versus practitioner: problems in social action research. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 283–293. (Originally published: *Social Work*, 1967, 12, No. 4, 89–96).

5. Benedict, B. A., *et al*. The clinical-experimental approach to assessing organizational change efforts. *Journal of Applied Behavioral Science*, 1967, 3, No. 3, 347–380.

6. Bernstein, I. N. & Freeman, H. E. *Academic and entrepreneurial research*. New York: Russell Sage Foundation, 1975.

7. Borich, G. D. (Ed.) *Evaluating educational programs and products*. Englewood Cliffs, N.J.: Educational Technology, 1974.

8. Borich, G. D. & Drezek, S. F. Evaluating instructional transactions. In G. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, N.J.: Educational Technology, 1974.

9. Boyle, P. G. Criteria for program priorities. Paper presented at a conference of Extension Home Economists Program Leaders, Washington, D.C., November, 1972.

10. Campbell, D. T. Reforms as experiments. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 187–223. (Originally published: *American Psychologist*, 1969, 24, No. 4, 409–429.)

11. Campbell, D. T. & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. H. Hellmuth (Ed.), *Disadvantaged child*. New York: Brunner/Mazel, 1970, 3. Pp. 185–210.

12. Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage, (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. (Reprinted as *Experimental and quasi-experimental design for research*. Chicago: Rand McNally, 1966.)

13. Carter, W. E. A taxonomy of evaluation models: use of evaluation models in program evaluation, April, 1975. ED 109 244

14. Cicirelli, V. G. Project Head Start, a national evaluation: brief of the study. In D. G. Hayes (Ed.), *The Britannica review of American education*. Chicago: Encyclopedia Britannica, 1969. Pp. 1, 235–243.

15. Cicirelli, V. G. Transactional evaluation in a national study of Head Start. In R. M. Rippey (Ed.), *Studies in transactional evaluation*. Berkeley, Calif.: McCutchan, 1973.

16. Cohen, D. K. Politics and research: evaluation of social action programs in education. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, 1972. (Originally published: *Review of Educational Research*, 1970, 40, No. 2, 213–238.)

17. Costa, C. H. Cost utility: an aid to decision-making. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., February/March, 1973. ED 074 623

18. Cox, C. B. A design for evaluation: a case example. *Indiana Social Studies Quarterly*, Autumn 1971, 24, No. 2, 5–12.

19. Cronbach, L. J. Course improvement through evaluation. *Teachers College Record*, 1963, 64, 672–683.

20. Dobbert, M. L. & Dobbert, D. J. A general model for complete ethnographic evaluations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 1976. ED 121 812

21. Dressel, P. L. *Handbook of academic evaluation*. San Francisco: Jossey-Bass, 1976.

22. Eash, M. J. and associates. Traditional vs. emergent values: a curriculum evaluation in an expanding suburban community. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 1976.

23. Ebel, R. L. *Measuring educational achievement*. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

24. Edwards, K. J. Summative evaluation: some basic considerations. In G. D. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, N.J.: Educational Technology, 1974.

25. Elinson, J. Effectiveness of social action programs in health and welfare. In C. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. (Originally published: Ross Laboratories, Columbus, Ohio, 1967.)

26. Evans, J. W. Evaluating educational programs—are we getting anywhere? *Educational Researcher*, September 1974, 7–10.

27. Farland, R. W. *et al.* The study of extended opportunity programs and services in California's community colleges. Final report to the Board of Governors, California Community Colleges, 1976.

28. Findlay, D. C. Application of the CIPP evaluation model to a center with multiple program areas and levels. *Education Technology*, October 1971, 11, No. 10, 43–47.

29. Glasser, R. Evaluation of instruction and changing educational models. In M. C. Wittrock & D. E. Wiley (Eds.), *The evaluation of instruction: issues and problems*. New York: Holt, Rinehart and Winston, 1970. Pp. 70–86.

30. Glass, G. V. Educational product evaluation: a prototype format applied. *Educational Researcher*, 1972, 1, No. 1, 1–4.

31. Glass, G. V. The growth of evaluation methodology. Research paper no. 27, Laboratory of Educational Research. Boulder, Colo.: Univer. of Colorado, 1969.

32. Glennan, T. K., Jr. Evaluating federal manpower programs: notes and observations. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 174–186. (Originally published: Santa Monica, Calif.: Rand Corporation, 1969.)

33. Greenberg, B. G. Evaluation of social programs. *Review of the International Statistical Institute*, 1968, 36, 260–277.

34. Guba, E. G. The failure of educational evaluation. *Educational Technology*, 1969, 9, No. 5, 29–38. Also in C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 250–266.

35. Guba, E. G. Methodological strategies for educational change. Paper presented at the conference on Strategies for Educational Change, Washington, D.C., November, 1965.

36. Guba, E. G. Problems in writing the results of evaluation. *Journal of Research and Development in Education*, 1975, 8, No. 3, 42–54.

37. Guba, E. G. & Stufflebeam, D. L. *Evaluation: the process of stimulating, aiding and abetting insightful action*. Bloomington, Ind.: Indiana Univer., 1970. ED 055 733

38. Guttentag, M. Models and methods in evaluation research. *Journal for the Theory of Social Behavior*, 1971, 1, No. 1, 75–95. Abstract no. 06093, v. 51, doc. yr. 1974.

39. Hecht, A. R. Utility of the CIPP model for evaluating an established career program in a community college. Paper presented at the annual meeting of the American Educational Research Association, San Franscisco, Calif., April 1976. ED 120 203

40. Hoepfner, R. *et al. CSE elementary school evaluation kit: needs assessment*. Boston: Allyn and Bacon, Inc., 1972.

41. House, E. R. Justice in evaluation. In C. V. Glass (Ed.), *Evaluation studies: review annual*. Beverly Hills, Calif.: Sage Publishing Co., 1976. Pp. 75–100.

42. Houston, T. R. Behavioral science impact-effectiveness model. In P. Rossi & W. Williams (Eds.), *Evaluating social programs*. New York: Seminar Press, 1972.

43. Hunter, M. G. & Schooley, D. E. The synergistic evaluation model Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., February/March, 1973. ED 755 503

44. Hyman, H. H., Wright, C. R., & Hopkins, T. K. *Applications of methods of evaluation: four studies of the encampment for citizenship*. Los Angeles, Calif.: Univer, of California Press, 1962.

45. Katz, D. S. & Morgan, R. L. A holistic strategy for the formative evaluation of educational programs. In G. D Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, N. J.: Educational Technology, 1974. Pp. 210–231.

46. Kourilsky, M. An adversary model for educational evaluation. *Evaluation Comment*, 1974, 4, No. 2.

47. Levine, M. Scientific method and the adversary model. *American Psychologist*, September, 1974, 666–679.

48. Lucco, R. J. Conceptualizing evaluation strategy: an evaluation systems framework. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April, 1976. ED 124 578

49. MacDonald, B. Evaluation and the control of education. Unpublished paper, Centre for Applied Research in Education, University of East Anglia, Norwich, England, May, 1974.

50. Madow, W. Project head start, a national evaluation: methodological critique. In D. G. Hays (Ed.), *The Britannica review of American education*. Chicago: Encyclopedia Britannica, 1969, 1. Pp. 245–260.

51. Mann, J. *Changing human behavior*. New York: Scribner's, 1965.

52. Mann, J. Evaluating educational programs: a symposium. *The Urban Review*, 1969, 3, No. 4, 12–13.

53. McDill, E. L., McDill, M. S. & Sprehe, J. T. Strategies for *success in compensatory education*. Baltimore: Johns Hopkins, 1969.

54. Metfessel, N. S. & Michael, W. B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement*, 1967, 27, 931–943.

55. Nyre, G. F. A view from the top looking sideways: professional schools and professional development. Paper presented at the annual meeting of the Professional and Organizational Development Network in Higher Education, Warrenton, Va., October 1976.

56. Owens, T. R. Educational evaluation by adversary proceedings. In E. R. House (Ed.), *School evaluation: the politics and process*. Berkeley, Calif.: McCutchan, 1973.

57. Parlett, M. & Hamilton, D. Evaluation as illumination: a new approach to the study of innovatory programs. Unpublished manuscript, 1974.

58. Peper, J. B. An ontological model of evaluation: a dynamic model for aiding organizational development. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., February/March, 1973. ED 078 039

59. Phi Delta Kappa, National Study Committee on Evaluation. *Educational evaluation and decision making*. Itasca, Ill.: Peacock Press, 1971.

60. Popham, W. J. *Educational evaluation*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1975.

61. Popham, W. J. (Ed.) *Evaluation in education*. Berkeley, Calif.: McCutchan, 1974.

62. Popham, W. J. & Carlson, D. Deep dark deficits of the adversary evaluation model. *Educational Researcher*, June, 1977, 6, No. 6, 3–6.

63. Porter, A. C. Analysis strategies for some common evaluation paradigms. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., February/March, 1973.

64. Provus, M. Evaluation of ongoing programs in the public school system. In B. R. Worthen & J. R. Sanders, (Eds.), *Educational evaluation: theory and practice*. Worthington, Ohio: Charles A. Jones, 1973.

65. Rhine, W. R. Strategies for evaluating Follow Through. In R. M. Rippey, (Ed.), *Transactional Evaluation*. Berkeley, Calif.: McCutchan, 1973.

66. Riecken, H. W. Memorandum on program evaluation. In C. H. Weiss, (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 85–104.

67. Rippey, R. M. (Ed.) *Studies in transactional evaluation*. Berkeley, Calif.: McCutchan, 1973.

68. Rose, C. & Nyre, G. F. *Access and assistance: the study of EOP/EOPS in California's public institutions of higher education. Volume I: analysis and recommendations*. Final report to the California Postsecondary Education Commission, 1976.

69. Rose, C. & Nyre, G. F. *Access and assistance: the study of EOP/EOPS in California's public institutions of higher education. Volume II: case study profiles*. Final report to the California Postsecondary Education Commission, 1976.

70. Rose, C. & Nyre, G. F. *An evaluation of the Los Angeles Community College District's Instructional Development Grant Program*. Final report to the Office of the Chancellor, Los Angeles Community College District, 1975.

71. Rose, C. & Nyre, G. F. Holistic evaluation: an eclectic approach to program evaluation. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

72. Rossi, P. H. Boobytraps and pitfalls in the evaluation of social action programs. In C. H. Weiss, (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 224–235. (Originally published: Washington, D.C.: American Statistical Association, 1966, 127–132.)

73. Sax, G. *Principles of educational measurement and evaluation*. Belmont, Calif.: Wadsworth, 1974.

74. Scriven, M. Evaluation perspectives and procedures. In W. J. Popham (Ed.), *Evaluation in education: current applications*. Berkeley, Calif.: McCutchan, 1974. Pp. 3–93.

75. Scriven, M. Goal-free evaluation. In R. E. House (Ed.), *School evaluation: the politics and process*. Berkeley, Calif.: McCutchan, 1973.

76. Scriven, M. The methodology of evaluation. In R. E. Stake, (Ed.), *Perspectives of curriculum evaluation*. AERA Monograph Series on Curriculum Evaluation, no. 1. Chicago: Rand McNally, 1967.

77. Scriven, M. Prose and cons about goal-free evaluation. *Evaluation Comment*, December 1972, 3, No. 4.

78. Simmel, G. *The sociology of Georg Simmel.* Translated by K. H. Wolff. New York: The Free Press, 1964.

79. Smith, L. M. & Pohland, P. A. Education, technology and the rural highlands. In R. E. Stake, (Ed.), *Four evaluation examples: anthropology, economic, narrative and portrayal*. AERA Monograph Series on Curriculum Evaluation, no. 7. Chicago: Rand McNally, 1974.

80. Smith, M. S. & Bissell, J. S. Report analysis: the impact of Head Start. *Harvard Education Review*, 1970, 40, No. 1, 51–104.

81. Spindler, G. D. *The transmission of American culture*. Cambridge, Mass.: Harvard Univer. Press, 1962.

82. Stake, R. E. *The case study method in social inquiry*. Urbana, Ill.: Center for Instructional Research and Curriculum Evaluation, Univer. of Illinois, 1976.

83. Stake, R. E. The countenance of educational evaluation. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 31–51. (Originally published: *Teachers College Record*, April 1967, 68, No. 7, 523–540).

84. Stake, R. E. Program evaluation, particularly responsive evaluation. Occasional paper #5. Kalamazoo, Mich.: Evaluation Center, Western Michigan Univer., November 1975.

85. Stake, R. E. Toward a technology for the evaluation of educational programs. In Tyler, Ralph W., Gagne Robert M., and Scriven, Michael (Eds.), *Perspectives of curriculum evaluation*. AERA Monograph Series on Curriculum Evaluation, no. 1. Chicago: Rand McNally, 1967.

86. Stanley, J. C. Controlled field experiments as a model for evaluation. In P. Rossi & W. Williams (Eds.), *Evaluating social programs*. New York: Seminar Press, 1972.

87. Stanley, J. C. Reactions to the March article on significant differences. *Educational Researcher*, 1969, 20, No. 5, 8–9.

88. Stufflebeam, D. L. Alternative approaches to educational evaluation. In W. J. Popham (Ed.), *Evaluation in education: current applications*. Berkeley, Calif.: McCutchan, 1974.

89. Stufflebeam, D. L. The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education*, 1971, 5, No. 1, 19–25.

90. Stufflebeam, D. L. The use of experimental design in education. *Journal of Educational Measurement*, Winter 1971, 8, No. 4, 267–74.

91. Suchman, E. A. Action for what? A critique of evaluative research. In R. O'Toole (Ed.), *The organization, management and tactics of social research*. Cambridge, Mass.: Schenkman Publishing Co., Inc., 1970.

92. Suchman, E. A. *Evaluative research*. New York: Russell Sage Foundation, 1967.

93. Thorndike, R. L. & Hagen, E. *Measurement and evaluation in psychology and education*. New York: John Wiley, 1969.

94. Tripodi, T., Fellin, P., & Epstein, I. *Social program evaluation: guidelines for health, education and welfare administration*. Itasca, Ill.: F. E. Peacock, Inc., 1971.

95. Trow, M. Methodological problems in the evaluation of innovation. In M. C. Wittrock & D. E. Wiley (Eds.), *The evaluation of instruction: issues and problems*. New York: Holt, Rinehart and Winston, Inc., 1970.

96. Tyler, R. W. General statement on evaluation. *Journal of Educational Research*, 1942, 35, 492–501.

97. Webb, E. J. *et al. Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.

98. Weiss, C. H. *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972.

99. Weiss, C. H. *Evaluation research: methods of assessing program effectiveness*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

100. Weiss, C. H. The politicization of evaluation research. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. Pp. 327–338. (Originally published: *Journal of Social Issues*, 1970, 26, No. 4, 57–68.)

101. Weiss, R. S. & Rein, M. The evaluation of broad-aim programs: difficulties in experimental design and an alternative. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972, 236–249.

102. Welch, W. & Walberg, H. A national experiment in curriculum evaluation. *American Educational Research Journal*, 1972, 9, 373–384.

103. Wergin, J. J. Evaluating faculty development programs, 1976. (unpublished mimeographed paper)

104. Westinghouse Learning Corporation and Ohio University. The impact of Head Start: an evaluation of the effects of Head Start on children's cognitive and affective development. Springfield, Va.: U.S. Department of Commerce, Clearinghouse for Federal Scientific and Technical Information, 1969.

105. Wetherill, R. G. & Buttram, J. L. Alternative modes of evaluation and their application to rural development. Paper presented at the Rural Sociology Section of the SAAS Meetings, Mobile, Alabama, 1976. ED 121 557

106. White, S. H. The national impact study of Head Start. In J. H. Hellmuth (Ed.), *Disadvantaged child*. New York: Brunner/Mazel, 1970. Pp. 3, 163–184.

107. Wholey, J. S. *et al.* Proper organizational relationships. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education*. Boston: Allyn and Bacon, Inc., 1972. (Originally published: Federal evaluation policy: an overview, a summary of the Urban Institute Study of Social Program Evaluation by federal agencies, September 1969).

108. Wilson, S. *et al. The use of ethnography in educational evaluation*. Chicago: Center for New Schools, July 1974. ED 126 147

109. Wolf, R. L. Trial by jury: a new evaluation method. *Phi Delta Kappan*, November 1975, 57, No. 3, 185–187.

110. Wolf, R. L., Potter, J., & Baxter, B. The judicial approach to educational evaluation. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April, 1976.

111. Worthen, B. R. & Sanders, J. R. *Educational evaluation: theory and practice*. Worthington, Ohio: Charles A. Jones, 1973.

112. Yost, M. & Monnin, F. J. A system approach to the development of an evaluation system of ESEA Title III projects. Unpublished research report. ED 047 356